Australian National University

Asymptotics of Maximum Composite Likelihood Estimation for Geostatistical Data

Nelson Jinn-Yih Chua

Supervised by Prof A. H. Welsh and

Dr Francis K. C. Hui

A thesis submitted in partial fulfilment of the requirements
for the degree of Bachelor of Statistics with Honours in Statistics
at the Australian National University.

October 2018

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge and belief, contains no material published or written by another person, except where due reference is made in the thesis.

Nelson Jinn-Yih Chua

October 31, 2018

Acknowledgements

First and foremost, I would like to express my utmost gratitude for my supervisors Alan and Francis, who were absolutely pivotal to the development of this thesis. Alan's wealth of knowledge and experience ensured that I progressed with a high level of efficiency. Meanwhile, Francis provided extensive guidance on the application of the expanding domain and infill frameworks in both a geostatistical context and for thesis writing. Francis and I also had enjoyable pseudo-intellectual discussions about Convergence and partial residuals; albeit at the expense of Alan. Their tremendous dedication and passion is <v e r y g o o d> and much appreciated.

I would also like to thank my friends and family for their continued care and support over the year. Special thanks goes to Jiabin for enlightening me about the existence of parallel processing in `R`, without which I would still be hoping for my simulation to be complete by the end of next year. I would also like to thank Nickson for peer-reviewing my work throughout the year, as well as for engaging in pseudo-intellectual discussions with me.

Abstract

Parameter estimation and inference in a geostatistical model is often made challenging due to the strong dependence between nearby observations. For large sample sizes, maximum likelihood estimation quickly becomes computationally expensive to perform, so other estimation approaches such as maximum composite likelihood estimation have been proposed as alternatives. In this thesis, we investigate the statistical and computational performance of maximum composite likelihood estimation relative to maximum likelihood estimation for the Gaussian exponential covariance model. As the main contribution of this work, we derive and analyse the exact closed-form expressions for the sandwich covariance matrix of various composite likelihoods in one-dimensional space. These new results are found under a hybrid asymptotic framework, which unifies the traditional expanding domain and infill frameworks seen in the geostatistical literature. We then demonstrate the practical implementation of maximum composite likelihood approaches for estimation and inference, as well as perform a data-motivated simulation study of their statistical performance in a two-dimensional setting with irregularly-spaced observations.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Spatial data consists of observations that are collected over a geographical area. Such data are common in a wide variety of disciplines including ecology (Fortin et al., 2012), climatology (Nowak et al., 2017), demography (Matthews and Parker, 2013) and geology (Angelini and Heuvelink, 2018). The rise of the information age has seen a sharp increase in the amount of spatial data being collected; and what comes with it is a greater demand for analytical tools and methodologies to understand the data. Due to the size and complexity of these datasets, it is also important to consider methods of analysis and inference that are computationally efficient.

There are three main categories that spatial data fall under: geostatistical, areal (or lattice) and point process data (Cressie and Wikle, 2011, p. 124). Geostatistical data is where the variables of interest are observed at fixed collection points. On the other hand, areal data are concerned with variables that are aggregated over well-defined geographical areas, such as cities and territories. The implication of this in terms of asymptotics is that in a geostatistical setting, we are free to obtain more observations at different locations in our spatial domain, but for areal data, further observations are only possible as part of a longitudinal study; that is, through the introduction of a temporal dimension. Finally, point process data occur when the locations at which observations occur are random and may themselves be of interest. The methods and models used to analyse each of these types of spatial data are quite different, and we will focus our attention towards geostatistical data for this thesis. Our motivating data for this will be maximum temperature data of the United States, which has been recorded at over one thousand land surface stations by the National Oceanic and Atmospheric Administration (Peterson and Vose, 1997).

1.1 Geostatistical Framework

Consider observing a variable of interest z at a set of locations $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}$, where $\mathbf{s}_i \in \mathcal{S} \subseteq \mathbb{R}^d$ for $d \in \mathbb{Z}^+$. It is common in the literature to assume that these observations are subject to random additive measurement errors $\varepsilon(\mathbf{s}_i)$ (Cressie and Wikle, 2011, p. 121). Thus, z can be related to the true unobserved spatial process y according to the following:

$$z(\mathbf{s}_i) = y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i).$$

By imposing specific distributional assumptions on the processes $\{y(\mathbf{s})\}$ and $\{\varepsilon(\mathbf{s})\}$, this becomes a parametric geostatistical model. In particular, this thesis will focus on the case where $\{y(\mathbf{s})\}$ is a Gaussian process, as defined below, and $\varepsilon(\mathbf{s})$ are independent and identically distributed (i.i.d.) normal random variables with zero mean and variance τ^2, where τ^2 is known as the nugget effect (Cressie and Wikle, 2011, p. 121). By extension, due to the additive property of Gaussian random variables, this means that $\{z(\mathbf{s})\}$ is also a Gaussian process.

Definition 1.1 (*Gaussian process*) A process $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^d\}$ is Gaussian if its finite dimensional distributions are Gaussian; that is, for any finite subset of locations $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\} \subset \mathcal{S}$, we have that the random vector $(w(\mathbf{s}_1), w(\mathbf{s}_2), ..., w(\mathbf{s}_n))^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$.

A key feature of spatial data is the strong dependence between nearby observations. As per Tobler's first law of geography, "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970) Hence, it is important to account for spatial correlation when analysing spatial data. In a Gaussian process, this is captured exclusively through the covariance matrix $\boldsymbol{\Sigma}$.

1.2 Covariance and Variograms

A common simplifying assumption to make about a spatial process is for it to be (weakly) stationary (Cressie and Wikle, 2011, p. 129).

Definition 1.2 (*Stationary spatial process*) A spatial process $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^d\}$ is (weakly) stationary if $\mathbb{E}[z(\mathbf{s})] \equiv \mu_z$, $\text{var}[z(\mathbf{s})] < \infty$ and $\text{cov}[z(\mathbf{s}), z(\mathbf{s}+\mathbf{h})] = \mathbb{E}[(z(\mathbf{s}) - \mu_z)(z(\mathbf{s}+\mathbf{h}) - \mu_z)] \equiv C_z(\mathbf{h})$ for all $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{S}$.

Under the assumption of stationarity, the strength of dependence between observations, as expressed through the covariance, is only dependent on their distance and direction apart. This acts as a kind of smoothness which aids in the interpretability of a parametric model by reducing the number of parameters needed to describe a system. It is often used when the behaviour of the response variable is thought to be roughly homogeneous over the space S. If, further to this, it is believed that the dependence decays in a radial manner from each location, then an additional restriction of covariance depending only on the some metric $\|\cdot\|$ (such as Euclidean distance) can be imposed:

Definition 1.3 (*Isotropic covariance*) A covariance structure is isotropic if $\text{cov}[z(\mathbf{s}), z(\mathbf{s}+\mathbf{h})] \equiv C_z(\|\mathbf{h}\|)$ for all $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{S}$.

In the time series literature, a common tool used to select an appropriate model for the covariance structure is the autocorrelation function, which estimates the covariance between observations at each time difference. However, due to the often multidimensional nature of $\mathcal{S}$ and the presence of a nugget effect, it is common to use a variogram instead when analysing geostatistical data (Gneiting et al., 2001).

Definition 1.4 (*Variogram*) The variogram for a stationary covariance structure is calculated by $2\gamma_z(\mathbf{h}) \equiv \text{var}[z(\mathbf{s}) - z(\mathbf{s}+\mathbf{h})] = 2(C_z(\mathbf{0}) - C_z(\mathbf{h}))$. The quantity $\gamma_z(\mathbf{h})$ is known as the semivariogram.

For practical usage, an empirical semivariogram $\hat{\gamma}_z(\mathbf{h})$ is computed from mean-stationary data z. An omnidirectional semivariogram provides an estimate of $\hat{\gamma}_z(\mathbf{h})$ at a set of distances $\|\mathbf{h}\|$ by considering

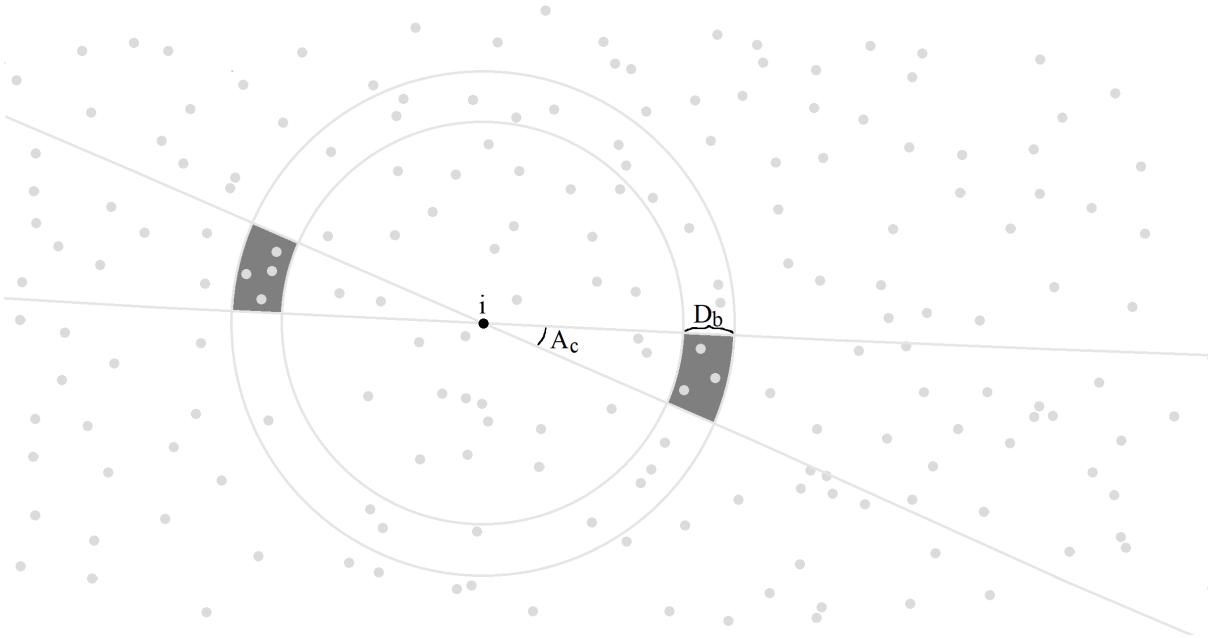

Figure 1.1: Region of leeway for a directional semivariogram. For all i, we identify the observations indexed by $j > i$ that lie in the tolerance region T_{bc} (filled dark grey regions).

the mean squared difference of z for all pairs of observations that are $\|\mathbf{h}\|$ units apart, with a degree of leeway. If, in addition, the leeway is further partitioned by direction, as illustrated in Figure 1.1, this is known as a directional semivariogram. This is typically used to assess the validity of assuming an isotropic covariance structure.

Definition 1.5 (*Empirical semivariogram for two-dimensional space*) Let $\mathcal{S} \in \mathbb{R}^2$ and consider a set of partitions with distance cut-offs $0 = d_0 < d_1 < ... < d_{max}$ and angle cut-offs $-\pi/2 - \delta = a_0 < a_1 < ... < a_{max} = \pi/2 - \delta$ with $0 \leq \delta < \pi$. Let the separation vector be denoted as $\mathbf{h}_{ij} = \mathbf{s}_i - \mathbf{s}_j = (h_{ij,1}, h_{ij,2})^T$ for all $1 \leq i < j \leq N$. For each pair of intervals (D_b, A_c), where $D_b = (d_b, d_{b+1}]$, $A_0 = (\pi/2 - \delta, \pi/2] \cup (-\pi/2, a_1]$ and $A_c = (a_c, a_{c+1}]$ for $c \neq 0$, let $T_{bc} = \{(i,j) : ||\mathbf{h}_{ij}|| \in D_b, \arctan(h_{ij,2}/h_{ij,1}) \in A_c\}$ be the region of leeway. Then the empirical semivariogram for the mid-point of D_b and A_c is given by

$$\hat{\gamma}_z(D_b, A_c) \equiv \frac{1}{2|T_{bc}|} \sum_{(i,j) \in T_{bc}} (z(\mathbf{s}_i) - z(\mathbf{s}_j))^2,$$

where $|\cdot|$ is the cardinality of the set. If $a_{max} = a_1$, then the empirical semivariogram is omnidirectional; otherwise, it is directional.

The shape of an empirical semivariogram with respect to distance is then assessed in order to determine an appropriate parametric model. Commonly used isotropic covariance structures include the spherical and Matérn; the latter of which encompasses the exponential and squared exponential as special cases (Stein, 1999, p. 31). For this thesis, we will focus on the exponential covariance structure as it is one of the simplest and widely explored cases in the literature (Gneiting et al., 2001).

Definition 1.6 (*Exponential covariance*) A mean-stationary spatial process $z(\mathbf{s})$ has an exponential covariance structure if $C_z(\mathbf{h}) = \tau^2 I(\|\mathbf{h}\| = 0) + \sigma^2 \exp(-\alpha\mathbf{h})$, or equivalently, $\gamma_z(\mathbf{h}) = \tau^2 I(\|\mathbf{h}\| \neq 0) + \sigma^2(1 - \exp(-\alpha\mathbf{h}))$, where $I(\cdot)$ is the indicator function.

Under this particular covariance structure, the covariance matrix $\boldsymbol{\Sigma}$ can be manipulated algebraically in certain basic geostatistical settings, such as having closed-form expressions for the inverse and determinant (Kac et al., 1953).

1.3 Maximum Likelihood Estimation

Once a parametric spatial model has been specified for the data $\mathbf{z} = (z(\mathbf{s}_1), z(\mathbf{s}_2), ..., z(\mathbf{s}_N))^T$, with unknown true parameter values stored in the vector $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ for $p \in \mathbb{Z}^+$ and $p \leq n$, it is of interest to estimate and perform inference on $\boldsymbol{\theta}$. One widely used starting point for this is maximum likelihood estimation, where an estimate $\hat{\boldsymbol{\theta}}$ is chosen such that it maximises the likelihood of obtaining the observed data.

Definition 1.7 (*Maximum likelihood estimate*) The maximum likelihood estimate of $\boldsymbol{\theta}_0$ under a specified model for $\mathbf{z}$ with joint density $f(\mathbf{z};\boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta};\mathbf{z})$ is $\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, \mathcal{L}(\boldsymbol{\theta};\mathbf{z})$. Equivalently, one can maximise the log-likelihood function; that is, $\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, \ell(\boldsymbol{\theta};\mathbf{z}) \equiv \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\, \log \mathcal{L}(\boldsymbol{\theta};\mathbf{z})$.

Such optimisation problems are usually solved by taking the first-order partial derivatives of the objective function with respect to $\boldsymbol{\theta}$, setting each resulting equation to zero, and solving the system of p equations. In the context of maximum likelihood estimation, the vector of first-order partial derivatives is known as

the score function $\text{sc}(\boldsymbol{\theta};\mathbf{z}) \equiv \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathbf{z})$. After solving $\text{sc}(\boldsymbol{\theta};\mathbf{z}) = \mathbf{0}$, the (observed) information function $\text{info}(\boldsymbol{\theta};\mathbf{z}) \equiv -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ell(\boldsymbol{\theta};\mathbf{z})$ may then be used to determine the nature of the stationary points found. In particular, a stationary point is (at least) a local maximum if the information matrix evaluated at the stationary point is positive definite.

The following theorem by Bartlett (1953) highlights a useful property of the score function:

Theorem 1.1 (*Expected score equals zero*) Suppose that the interchange of differentiation and integration is permissible at $\boldsymbol{\theta}_0 \in \Theta$. Then $\mathbb{E}[\text{sc}(\boldsymbol{\theta}_0;\mathbf{z})] = 0$.

Proof: $\mathbb{E}[\text{sc}(\boldsymbol{\theta}_0;\mathbf{z})] = \int_{-\infty}^{\infty}\text{sc}(\boldsymbol{\theta}_0;\mathbf{z})f(\mathbf{z};\boldsymbol{\theta}_0)\,d\mathbf{z} = \int_{-\infty}^{\infty}\frac{\frac{\partial}{\partial\boldsymbol{\theta}_0}f(\mathbf{z};\boldsymbol{\theta}_0)}{f(\mathbf{z};\boldsymbol{\theta}_0)}f(\mathbf{z};\boldsymbol{\theta}_0)\,d\mathbf{z} = \frac{\partial}{\partial\boldsymbol{\theta}_0}\int_{-\infty}^{\infty}f(\mathbf{z};\boldsymbol{\theta}_0)\,d\mathbf{z} = \frac{\partial}{\partial\boldsymbol{\theta}_0}(1) = 0$.

This is a necessary condition for the maximum likelihood estimator to be asymptotically unbiased (Lindsay, 1988).

A quantity associated with the variances and covariances of the estimator $\hat{\boldsymbol{\theta}}$ is the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}) \equiv \text{var}[\text{sc}(\boldsymbol{\theta};\mathbf{z})] = \mathbb{E}[\text{sc}(\boldsymbol{\theta};\mathbf{z})\text{sc}(\boldsymbol{\theta};\mathbf{z})^T]$. Alternatively, we can use Theorem 1.1 to show that the Fisher information matrix can also be computed using $\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}[\text{info}(\boldsymbol{\theta};\mathbf{z})]$ (Bartlett, 1953). The Fisher information matrix will be shown to play an important role in the asymptotic properties of the maximum likelihood estimator in Section 1.4.

1.4 Asymptotics of Maximum Likelihood Estimation

A desirable property of any estimator is consistency, meaning that it will approach the true parameter value $\boldsymbol{\theta}_0$ as more data are collected.

Definition 1.8 (*Consistency of an estimator*) An estimator $\tilde{\boldsymbol{\theta}}$ is (weakly) consistent if it converges in probability to the true parameter value: $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. More formally, for any $\varepsilon > 0$, it holds that $\mathbb{P}(||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| > \varepsilon) \to 0$ as the sample size $N \to \infty$.

We may alternatively consider the notion of consistency in mean square error, which is a sufficient condition for consistency (by Chebyshev's inequality) and usually easier to apply.

Definition 1.9 (*Consistency in mean square error*) An estimator $\tilde{\boldsymbol{\theta}}$ is consistent in mean square error if the mean square error $\text{mse}[\tilde{\boldsymbol{\theta}}] \equiv \mathbb{E}[(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2] = \{\text{bias}[\tilde{\boldsymbol{\theta}}]\}^2 + \text{var}[\tilde{\boldsymbol{\theta}}]$ satisfies $\lim_{N\to\infty} \text{mse}[\tilde{\boldsymbol{\theta}}] = 0$.

It is also desirable for an estimator to have low variability. Cramér (1946, p. 477-481) and Rao (1945), amongst other statisticians around the same time, derived the theoretical lowest variance attainable by an unbiased estimator.

Theorem 1.2 (*Cramér-Rao lower bound*) Suppose that $\tilde{\boldsymbol{\theta}}$ is an unbiased estimator for $\boldsymbol{\theta}_0$. Then $\text{var}[\tilde{\boldsymbol{\theta}}] \geq \mathbf{I}(\boldsymbol{\theta}_0)^{-1}$.

Note that the above theorem holds regardless of the dependence structure of the data. However, in the classical scenario in which asymptotics of the maximum likelihood estimator were established, the data are assumed to be independent and identically distributed. The following theorem highlights the key asymptotic result in this context; it is presented in many statistical inference textbooks such as Casella and Berger (2002).

Theorem 1.3 (*Maximum likelihood asymptotics for i.i.d. data*) Let $z_1, z_2, ..., z_N \overset{iid}{\sim} f(z; \boldsymbol{\theta}_0)$, where f satisfies various regularity conditions (including differentiability with respect to $\boldsymbol{\theta}$ and parameter identifiability; see Casella and Berger (2002, p. 516) for further details). Then the maximum likelihood estimator follows an asymptotically normal distribution; that is, $\hat{\boldsymbol{\theta}}_{\text{ML}} \dot{\sim} N(\boldsymbol{\theta}_0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1})$.

This theorem implies that the maximum likelihood estimator is consistent and asymptotically achieves the Cramér-Rao lower bound.

1.5 Maximum Composite Likelihood Estimation and Asymptotics

Maximum composite likelihood estimation involves deliberately using a misspecified but structurally simpler likelihood as the objective function to be maximised, in place of the full likelihood $\mathcal{L}(\boldsymbol{\theta};\mathbf{z})$. The motivation behind using this estimation approach for geostatistical models is due to the complications associated with a strong dependence structure. The full likelihood may be difficult to write down explicitly; or even if we can, maximising this function may be computationally intractable. As an example, in the case of a Gaussian process, evaluating the full likelihood relies on finding the determinant and inverse of the $N \times N$ covariance matrix $\boldsymbol{\Sigma}$, both of which have computational costs of $O(N^3)$ in the absence of any exploitable matrix structure. This is problematic from a computational standpoint if the number of observations in the dataset to be analysed is large.

Varin et al. (2011) classify composite likelihoods into two broad categories, which describe whether the likelihood is composed of marginal or conditional densities, respectively:

Definition 1.10 (*Types of composite likelihood*) Let $\mathcal{B}_k$ correspond to some subset of the observations $\{z(\mathbf{s}_1),...,z(\mathbf{s}_N)\}$. A composite likelihood $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z})$ is called marginal if it is of the form $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) = \prod_{b=1}^{B} f(\mathcal{B}_b;\boldsymbol{\theta})$, and conditional if it is of the form $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) = \prod_{i=1}^{N} f(y(\mathbf{s}_i)|\mathcal{B}_i;\boldsymbol{\theta})$.

Once a composite likelihood has been constructed, the maximum composite likelihood estimator $\hat{\boldsymbol{\theta}}_{CL}$ can be found in a similar manner to maximum likelihood estimation. Firstly, we can find stationary points of the composite log-likelihood $c\ell(\boldsymbol{\theta};\mathbf{z}) \equiv \log \mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z})$ using the composite score function $\mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} c\ell(\boldsymbol{\theta};\mathbf{z})$. We can then identify maxima using the composite information function $\mathrm{info}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) \equiv -\frac{\partial}{\partial \boldsymbol{\theta}^T} \mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z})$.

The asymptotics of maximum composite likelihood estimation differ slightly from maximum likelihood estimation. Firstly, we note that composite likelihood functions are almost always constructed to satisfy the necessary condition for asymptotically unbiased estimation (Lindsay, 1988), which has a similar

proof to Theorem 1.1:

Theorem 1.4 (*Expected composite score equals zero*) Suppose that the interchange of differentiation and integration is permissible at $\boldsymbol{\theta}_0$. Then $\mathbb{E}[\text{sc}_\text{C}(\boldsymbol{\theta}_0;\mathbf{z})] = 0$, where the expectation is taken with respect to the true likelihood of the data $f(\mathbf{z};\boldsymbol{\theta}_0)$.

Proof: Note that the conditional densities $f(z(\mathbf{s}_i)|\mathcal{B}_i;\boldsymbol{\theta})$ that comprise a composite conditional likelihood can be expressed as the ratio of two marginal densities. Hence, both the composite marginal log-likelihood and a composite conditional log-likelihood can be written as a linear combination of marginal log-densities; that is, $c\ell(\boldsymbol{\theta};\mathbf{z}) = \sum_b m_b \log f(\mathcal{B}_b;\boldsymbol{\theta})$, with coefficients $m_b \in \{-1,1\}$. Thus, $\mathbb{E}[\text{sc}_\text{C}(\boldsymbol{\theta}_0;\mathbf{z})] = \sum_b \mathbb{E}[\frac{\frac{\partial}{\partial\boldsymbol{\theta}_0} f(\mathcal{B}_b;\boldsymbol{\theta}_0)}{f(\mathcal{B}_b;\boldsymbol{\theta}_0)}] = \sum_b \int_{-\infty}^{\infty} \frac{\partial}{\partial\boldsymbol{\theta}_0} f(\mathcal{B}_b;\boldsymbol{\theta}_0)\, d\mathcal{B}_b = 0$.

However, the asymptotic variance of maximum composite likelihood estimation is a quantity $\mathbf{G}(\boldsymbol{\theta})^{-1}$ known as the sandwich covariance matrix, which takes the place of the inverse Fisher information $\mathbf{I}(\boldsymbol{\theta})^{-1}$. This is highlighted in the following theorem available from Kent (1982) and Lindsay (1988):

Theorem 1.5 (*Maximum composite likelihood asymptotics for i.i.d. data*) Let $z_1,...,z_N \overset{iid}{\sim} f(z;\boldsymbol{\theta}_0)$, where the densities comprising the composite likelihood satisfy the regularity conditions from Theorem 1.3. Define $\mathbf{J}(\boldsymbol{\theta}) \equiv \text{var}[\text{sc}_\text{C}(\boldsymbol{\theta};\mathbf{z})] = \mathbb{E}[\text{sc}_\text{C}(\boldsymbol{\theta};\mathbf{z})\text{sc}_\text{C}(\boldsymbol{\theta};\mathbf{z})^T]$ and $\mathbf{H}(\boldsymbol{\theta}) \equiv \mathbb{E}[\text{info}_\text{C}(\boldsymbol{\theta};\mathbf{z})]$, which comprise the sandwich information matrix $\mathbf{G}(\boldsymbol{\theta}) \equiv \mathbf{H}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{H}(\boldsymbol{\theta})$. Then the asymptotic distribution of the maximum composite likelihood estimator is given by $\hat{\boldsymbol{\theta}}_\text{CL} \dot{\sim} N(\boldsymbol{\theta}_0, \mathbf{G}(\boldsymbol{\theta}_0)^{-1})$.

Note that in the case of the full (correctly specified) likelihood function, we have that $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})$, which collapses down to maximum likelihood estimation and Theorem 1.3.

It is important to note that the use of the structurally simpler composite likelihood over the full likelihood involves a trade-off between computational efficiency and statistical efficiency. In particular, it is often the case that $\mathbf{I}(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})$ is a positive semi-definite matrix (Varin, 2008), so there is less information from the data being utilised in estimating $\boldsymbol{\theta}_0$ using $\hat{\boldsymbol{\theta}}_\text{CL}$ than $\hat{\boldsymbol{\theta}}_\text{ML}$. Thus, we would like to investigate the extent of information loss from various choices of composite likelihood, and will be the focus of this thesis. This is often measured by computing the asymptotic relative efficiency, where we take the ratio

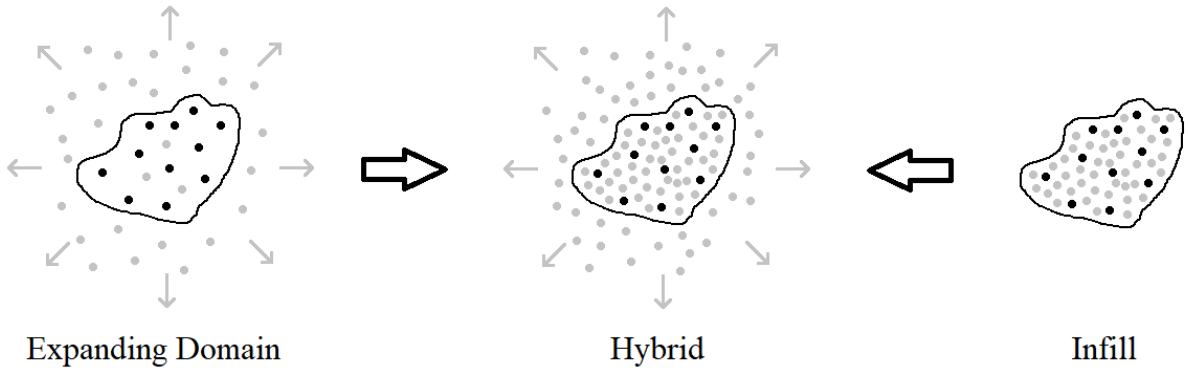

Figure 1.2: Comparison of asymptotic frameworks for geostatistical data. The black-bordered region and points denote the initial region of collection and locations of observations; grey points are further observation locations under the corresponding asymptotic framework.

of the elements in $\mathbf{I}(\boldsymbol{\theta})^{-1}$ to the corresponding elements in $\mathbf{G}(\boldsymbol{\theta})^{-1}$.

1.6 Asymptotics in a Geostatistical Framework

The classical asymptotic results for maximum likelihood estimation and maximum composite likelihood estimation, as highlighted in Theorems 1.3 and 1.5, rely on the assumption that the data are independent and identically distributed. However, since this does not hold in the context of geostatistical models, there is no guarantee that the favourable asymptotic properties of these two approaches still hold. Additionally, as illustrated in Figure 1.2, there are generally three different asymptotic frameworks that can be considered for geostatistical data: expanding domain, infill and hybrid.

Expanding domain asymptotics assume that observations continue to be taken over an increasing region of space out to infinity. This can be likened to taking future observations in a time series, where time is treated as a uni-directional one-dimensional space. Infill asymptotics, on the other hand, focus on taking an increasing density of observations within a closed region. It is often the case in spatial data that the geographical region of interest is closed, so analysis of infill asymptotics is desired.

By combining the notions of expanding domain and infill together, a third framework called hybrid asymptotics can also be obtained. The unification of these two frameworks allows us to take advantage

of the favourable asymptotic properties of the expanding domain framework (see Section 2.1) and use them in an infill context, including spatial prediction. This has applications in contexts where we can increase both the spatial resolution and domain of our observations, such as demographic studies (Lu and Tjøstheim, 2014).

Although hybrid asymptotics have been considered in papers as early as Hall and Patil (1994) and Lahiri et al. (2002), the framework has largely been overlooked until more recently. A unifying formalisation of all three asymptotic frameworks is presented in Lu and Tjøstheim (2014), where they define $\Delta_{j,N} \equiv \max\{||\mathbf{s}_i - \mathbf{s}_j|| : 1 \leq i \leq N, i \neq j\}$ and $\delta_{j,N} \equiv \min\{||\mathbf{s}_i - \mathbf{s}_j|| : 1 \leq i \leq N, i \neq j\}$, which correspond to the maximum and minimum distance between $\mathbf{s}_j$ and the other locations in the set of size N, respectively. These quantities are used in the following definition:

Definition 1.11 (*Asymptotic frameworks*) Consider a sequence of subsets of locations $S_1, S_2, ..., S_N \subset \mathcal{S}$, where S_j has j locations and the sequence is not necessarily nested. Then the following conditions define the corresponding asymptotic frameworks:

- Expanding domain: $\Delta_N \equiv \min_{1 \leq j \leq N} \Delta_{j,N} \to \infty$ as $N \to \infty$ and $\lim_{N \to \infty} \min_{1 \leq i \leq N} \delta_{i,N} \geq L > 0$
- Infill: $\delta_N \equiv \max_{1 \leq j \leq N} \delta_{j,N} \to 0$ as $N \to \infty$ and $\lim_{N \to \infty} \max_{1 \leq i \leq N} \Delta_{i,N} \leq U < \infty$
- Hybrid: $\Delta_N \to \infty$ and $\delta_N \to 0$

In this thesis, the sequence of sampling locations will be structured in a way that will allow for analysis of spatial models under the hybrid framework. This will also allow for analysis under the expanding domain and infill frameworks.

1.7 Thesis Outline

The primary objective of this thesis is to compare the statistical and computational performance of maximum composite likelihood estimation relative to maximum likelihood estimation in a geostatistical setting. It is not expected that maximum composite likelihood estimation will be more efficient than

maximum likelihood estimation, but it is important to discern whether using a misspecified likelihood is a viable alternative when the sample size is large and using the full likelihood is no longer feasible.

In Chapter 2, we shall review the literature on maximum likelihood and maximum composite likelihood estimation in a Gaussian geostatistical setting. In particular, we are interested in the asymptotic frameworks from Definition 1.11 where asymptotic normality of these estimators is shown to still hold even without the i.i.d. data assumption. We will then narrow our focus to results for maximum likelihood estimation under a Gaussian exponential covariance model, which has been widely studied in the literature. We will also highlight some choices of composite likelihood that have been studied, which reveals the lack of theoretical asymptotic results in the literature.

As the main contribution of this thesis, in Chapter 3, we explore the theoretical asymptotic efficiency of two choices of composite likelihood relative to the full likelihood in a one-dimensional exponential covariance model. We derive exact closed-form expressions for the sandwich covariance matrix $\mathbf{G}(\theta)^{-1}$ for these two choices and compare this to the inverse Fisher information $\mathbf{I}(\theta)^{-1}$ to obtain the asymptotic relative efficiency of these maximum composite likelihood estimators. This leads to new insights that would have been difficult to uncover from a simulation study alone, such as the effect of the strength of spatial correlation on efficiency.

In Chapter 4, we look at the practical implementation of composite likelihoods in the context of modelling maximum temperature data for the United States. This will involve the derivation of a broadly applicable variance estimator for maximum composite likelihood estimates in a Gaussian setting. We will use this example to motivate a simulation study in the two-dimensional exponential covariance setting, and draw comparisons with our findings in Chapter 3.

Finally, we will summarise our findings and contributions in Chapter 5. We will also identify potential future research directions that can further the developments in this thesis.

Chapter 2

Literature Review

Classical asymptotics to do with maximum likelihood and maximum composite likelihood estimation are derived under the assumption that the data are independent and identically distributed. Hence, in Section 2.1 of this literature review, we will highlight some of the asymptotic results available for maximum likelihood estimators in a geostatistical setting. In particular, we are interested in the conditions that are sufficient for consistency and asymptotic normality to still hold. Close attention will be paid to the asymptotic framework that is being considered, and how it is compatible with Definition 1.11.

Next, in Section 2.2, we will review the results that are available for maximum likelihood estimation asymptotics under a Gaussian exponential covariance model. This will cover many different cases, such as the dimension of space and whether or not a nugget effect is included. We will check that these results are consistent with the general theory in Section 2.1. The findings here will also allow us to set up a reliable benchmark when comparing maximum composite likelihood estimation to maximum likelihood estimation.

In Section 2.3, we will highlight the various choices of composite likelihood that have been previously studied in the geostatistical literature. For each type of composite likelihood, we will look at the motivation behind its construction, as well as any theoretical or numerical studies of their statistical performance. A few of these composite likelihood functions will be explored in our exponential covariance case during the later chapters.

2.1 Maximum Likelihood Estimation in Geostatistics

One of the earliest papers dealing with maximum likelihood asymptotics in a geostatistical setting is due to Mardia and Marshall (1984). Here, they worked in the context of Gaussian spatial regression, where $\mathbf{z} = (z(\mathbf{s}_1), z(\mathbf{s}_2), ..., z(\mathbf{s}_N))^T \sim N(\mathbf{X}\boldsymbol{\beta}_0, \boldsymbol{\Sigma}(\boldsymbol{\phi}_0))$ for an $N \times q$ covariate matrix $\mathbf{X}$ and vector of true parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T)^T \in \Theta \subseteq \mathbb{R}^{q+p}$. Applying some general results from Sweeting (1980), they identified sufficient conditions for the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{ML}}$. These conditions pertain to the continuity, growth and convergence of the observed information function $\text{info}(\boldsymbol{\theta}; \mathbf{z}) \equiv -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell(\boldsymbol{\theta}; \mathbf{z})$ and Fisher information $\mathbf{I}(\boldsymbol{\theta}) \equiv \text{var}[\text{sc}(\boldsymbol{\theta}; \mathbf{z})] = \mathbb{E}[\text{sc}(\boldsymbol{\theta}; \mathbf{z})\text{sc}(\boldsymbol{\theta}; \mathbf{z})^T]$.

> **Theorem 2.1** (*Maximum likelihood asymptotics for spatial regression*) Suppose that the usual regularity conditions from Theorem 1.3 hold, in addition to the covariance function having continuous second-order partial derivatives with respect to $\boldsymbol{\phi}$. Furthermore, assume $\lim_{N\to\infty} \mathbf{I}(\theta)^{-1} = 0$ and that $-\mathbf{I}(\boldsymbol{\theta})^{-\frac{1}{2}} \text{info}(\boldsymbol{\theta}) \mathbf{I}(\boldsymbol{\theta})^{-\frac{1}{2}}$ converges in probability to the identity matrix. Then $\hat{\boldsymbol{\theta}}_{\text{ML}} \dot{\sim} N(\boldsymbol{\theta}_0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1})$ with a convergence rate of $\sqrt{N}$.

In practice, Theorem 2.1 is often difficult to verify, so Mardia and Marshall (1984) then considered sufficient conditions for asymptotic normality to hold under the expanding domain framework.

> **Theorem 2.2** (*Maximum likelihood expanding domain asymptotics*) Let $\mathbf{z}$ have a stationary covariance structure $C_z(\mathbf{h})$, and $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}$ form an equally-spaced regular lattice on $\mathbb{R}^d$. Define $\mathcal{H}_N \equiv \{\mathbf{s}_i - \mathbf{s}_j; i, j \in \{1, 2, ..., N\}\}$ to be a set containing all of the unique displacements between any two observations on the lattice. Suppose that $\lim_{N\to\infty} (\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{0}$ and the parameters in $\boldsymbol{\phi}$ are not asymptotically linearly dependent. If $\lim_{N\to\infty} \sum_{\mathbf{h}\in\mathcal{H}_N} |C_z(\mathbf{h})| < \infty$, $\lim_{N\to\infty} \sum_{\mathbf{h}\in\mathcal{H}_N} |\frac{\partial}{\partial \phi_k} C_z(\mathbf{h})| < \infty$ and $\lim_{N\to\infty} \sum_{\mathbf{h}\in\mathcal{H}_N} |\frac{\partial^2}{\partial \phi_k \partial \phi_l} C_z(\mathbf{h})| < \infty$ for all $k, l = 1, 2, ..., p$, then we have that $\hat{\boldsymbol{\theta}}_{\text{ML}} \dot{\sim} N(\boldsymbol{\theta}_0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1})$ with a convergence rate of $\sqrt{N}$.

The above theorem is only applicable in the expanding domain framework due to the requirement for the covariance function and its derivatives to be absolutely summable. In essence, this ensures that there is

enough information about the covariance parameters ϕ available in the data, and this is reliant on having some observations that are far enough apart such that they are effectively uncorrelated. Due to the nature of an expanding domain, as the sample size N increases, displacements $\mathbf{h}$ that are progressively introduced to the set $\mathcal{H}_N$ will get larger in magnitude, and thus contribute comparably less to these sums. In contrast, under the infill framework, new displacements that are added to $\mathcal{H}_N$ will be near $\mathbf{0}$, causing the sums to diverge. This suggests that observations that are close to each other are too strongly correlated and hence provide less information about ϕ, which in turn demonstrates that the asymptotic behaviour of the maximum likelihood estimator under different asymptotic frameworks can vary.

A consequence of Theorem 2.2 is that if ϕ involves a nugget effect term τ^2, then asymptotic normality of $\boldsymbol{\theta}$ holds if and only if it would otherwise hold in the absence of the nugget effect; that is, for the latent spatial process $\mathbf{y}$. Since the nugget effect only appears as an additive term in the covariance function at a displacement of $\mathbf{0}$, we have that $\sum_{\mathbf{h}\in\mathcal{H}_N}|C_z(\mathbf{h})| = \tau^2 + \sum_{\mathbf{h}\in\mathcal{H}_N}|C_y(\mathbf{h})|$ and $\sum_{\mathbf{h}\in\mathcal{H}_N}|\frac{\partial}{\partial\tau^2}C_z(\mathbf{h})| = 1$, with no other differences between the sums for $\mathbf{z}$ and $\mathbf{y}$.

In the context of a spatial autoregressive model, Zheng and Zhu (2012) obtained general results for maximum likelihood asymptotics under the expanding domain, infill and hybrid frameworks. This extends on the results of Lee (2004) who considered the expanding domain framework alone. The spatial autoregressive model is given by $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where the errors are modelled using the autoregressive structure $\boldsymbol{\epsilon} = \mathbf{W}_N(\phi_0)\boldsymbol{\epsilon} + \boldsymbol{\nu}$, and the weighting matrix $\mathbf{W}_N(\phi) = (w_{ij,N})_{N\times N}$ has a diagonal of zeroes and $\nu_i \overset{iid}{\sim} N(0,\eta_0^2)$.

An essential component of the asymptotic theory by Zheng and Zhu (2012) is the rate m_N at which the off-diagonal elements in the weighting matrix decay to zero; that is, $w_{ij,N} = O(m_N^{-1})$ for $i \neq j$. Based on this, $m_N = O(1)$ corresponds to expanding domain, $m_N \to \infty$ with $m_N/N \to C > 0$ corresponds to infill, and $m_N \to \infty$ with $m_N/N \to 0$ corresponds to the hybrid framework. As an example, we can consider a common choice for weights which uses distance-based neighbours, where each row sums to one and equal weighting is assigned to observations within a certain distance of the observation under

consideration. Under expanding domain, the number of distance-based neighbours will remain finite so the weights will not decay to zero; but under the infill and the hybrid frameworks, the number of distance-based neighbours will grow to infinity. However, in the hybrid case, as long as the domain expands, N will grow at a faster rate than m_N, so that m_N/N will converge to zero.

Theorem 2.3 (*Maximum likelihood asymptotics for spatial autoregression*) Assume a set of regularity conditions related to the boundedness of $\mathbf{W}_N(\boldsymbol{\phi})$ and $\mathbf{X}$ are satisfied (see Zheng and Zhu (2012) for more details). Then $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ and $\hat{\eta}^2_{\mathrm{ML}}$ are consistent and follow an asymptotic normal distribution with a convergence rate of $\sqrt{N}$. However, the asymptotics for $\hat{\boldsymbol{\phi}}_{\mathrm{ML}}$ differ depending on the asymptotic framework being considered.

- If $m_N = O(1)$ then $\hat{\boldsymbol{\phi}}_{\mathrm{ML}} \xrightarrow{p} \boldsymbol{\phi}_0$. Furthermore, if $\lim_{N\to\infty} N^{-1}\mathbf{I}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ exists and is positive definite, where $\mathbf{I}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ is the block of the Fisher information matrix dealing exclusively with partial derivatives of $\boldsymbol{\phi}$, then $\hat{\boldsymbol{\phi}}_{\mathrm{ML}} \dot{\sim} N(\boldsymbol{\phi}_0, \mathbf{I}_{\boldsymbol{\phi}}(\boldsymbol{\theta}_0)^{-1})$ with a convergence rate of $\sqrt{N}$.
- If $m_N \to \infty$ and $m_N/N \to 0$, then $\hat{\boldsymbol{\phi}}_{\mathrm{ML}} \xrightarrow{p} \boldsymbol{\phi}_0$. However, the convergence rate is $\sqrt{N/m_N}$.
- If $m_N \to \infty$ and $m_N/N \to C > 0$, then the consistency of $\hat{\boldsymbol{\phi}}_{\mathrm{ML}}$ is not guaranteed.

Both Theorems 2.2 and 2.3 validate the idea that the usual asymptotics of maximum likelihood estimation can be translated to an expanding domain setting. We also see from Theorem 2.3 that asymptotic normality holds under a hybrid framework simply due the domain being able to expand. However, the asymptotics of maximum likelihood estimation for covariance parameters under the infill framework can be problematic. Once again, this highlights the idea that there is too little information about the strength of dependence between observations when the domain is fixed, even as the sample size increases.

Due to the favourable asymptotic behaviour of estimators under the expanding domain framework, most results in the geostatistical literature are derived in this setting. Meanwhile, infill is often only considered in specific cases such as the Gaussian exponential covariance model that will be discussed in Section 2.2. Finally, due to the lack of formalisation of the hybrid framework until more recently in Zheng and Zhu (2012) and Lu and Tjøstheim (2014), results in this situation are sparse.

2.2 Asymptotics for the Gaussian Exponential Covariance Model

As one of the most widely studied geostatistical models, we will now review some results for maximum likelihood estimation asymptotics under a Gaussian exponential covariance model, as per Definition 1.6. Given results such as Theorems 2.1 and 2.3, most of the interest in geostatistical asymptotics lies in the specified covariance structure rather than linear regression component, so this section will focus on zero-mean exponential covariance models.

In a one-dimensional setting on an equally-spaced lattice, Zhang and Zimmerman (2005) considered the expanding domain asymptotics of the Gaussian exponential covariance model. For this they used the observation locations $s_i = i$, and model $(z(s_0), z(s_1), ..., z(s_N))^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ with true parameter vector $\phi_0 = (\sigma_0^2, \alpha_0, \tau_0^2)^T$ and $\boldsymbol{\Sigma}_{0,ij} = \tau_0^2 I(i = j) + \sigma_0^2 \exp(-\alpha_0 |s_i - s_j|)$. They showed that in cases both with and without the nugget effect that the maximum likelihood estimator $\hat{\phi}_{\mathrm{ML}}$ is asymptotically normal under the expanding domain framework, and derived an explicit expression for the Fisher information matrix in both cases. In the nuggetless case (where $\tau_0^2 = 0$), they found that $\hat{\phi}_{\mathrm{ML}}$ is $\sqrt{N}$-consistent with asymptotic distribution

$$\sqrt{N}\begin{bmatrix} \hat{\sigma}^2_{\mathrm{ML}} - \sigma_0^2 \\ \hat{\alpha}_{\mathrm{ML}} - \alpha_0 \end{bmatrix} \dot{\sim} N\left(\mathbf{0}, \begin{bmatrix} 2(\sigma_0^2)^2 \frac{1+\mathrm{e}^{-2\alpha_0}}{1-\mathrm{e}^{-2\alpha_0}} & -2\sigma_0^2 \\ -2\sigma_0^2 & \mathrm{e}^{-2\alpha_0} - 1 \end{bmatrix}\right). \tag{2.1}$$

We will present a detailed derivation of the Fisher information matrix in the nuggetless case in Section 3.2 to verify (2.1), which will highlight some of the techniques and results required to derive the sandwich covariance matrix for composite likelihood functions.

Under infill, however, our classical asymptotics in the one-dimensional exponential covariance case do not hold. In the nuggetless case, with $0 \leq s_1 < s_2 < ... < s_N \leq 1$ and $\mathcal{S}_N = \{s_1, .., s_N\}$ not necessarily nested, Ibragimov and Rozanov (1978, p. 100) showed that asymptotically, σ^2 and α are indistinguishable for a given product $\alpha\sigma^2$. Based on this, Ying (1991) derived asymptotic distributions for $\hat{\alpha}_{\mathrm{ML}}\hat{\sigma}^2_{\mathrm{ML}}$,

as well as the maximum likelihood estimators for a single parameter when the other is fixed, which are as follows:

$$\sqrt{N}(\hat{\alpha}_{\mathrm{ML}}\hat{\sigma}^2_{\mathrm{ML}} - \alpha_0\sigma_0^2) \dot{\sim} N(0, 2\alpha_0\sigma_0^2),$$

$$\sqrt{N}(\underset{\sigma^2}{\operatorname{argmax}}\, \ell(\sigma^2, \alpha; \mathbf{y}) - \sigma_0^2) \dot{\sim} N(0, 2\sigma_0^4), \tag{2.2}$$

$$\sqrt{N}(\underset{\alpha}{\operatorname{argmax}}\, \ell(\sigma^2, \alpha; \mathbf{y}) - \alpha_0) \dot{\sim} N(0, 2\alpha_0^2). \tag{2.3}$$

The consequence of (2.2) and (2.3) is that there are situations where the asymptotic normality of maximum likelihood estimation still holds under the infill framework, but this is subject to stronger parameter identifiability conditions than in the expanding domain framework.

In the presence of a nugget effect, the asymptotic behaviour of our estimators can change under infill. Chen et al. (2000) showed that

$$\begin{bmatrix} \sqrt[4]{N}(\hat{\alpha}_{\mathrm{ML}}\hat{\sigma}^2_{\mathrm{ML}} - \alpha_0\sigma_0^2) \\ \sqrt{N}(\hat{\tau}^2_{\mathrm{ML}} - \tau_0^2) \end{bmatrix} \dot{\sim} N\left(\mathbf{0}, \begin{bmatrix} 4\sqrt{2}\tau_0(\alpha_0\sigma_0^2)^{\frac{3}{2}} & 0 \\ 0 & 2\tau_0^4 \end{bmatrix}\right).$$

Thus, it is also possible under infill to obtain non-standard asymptotic distributions, or ones with a slower rate of convergence. This contrasts to Theorem 2.2, where the usual asymptotics hold regardless of the presence of a nugget effect under the expanding domain framework.

Ying (1993) explored the infill asymptotics of the exponential covariance structure in a two-dimensional case. For this, they considered a sampling scheme on a rectangular lattice with locations $\mathbf{s}_{ij} = (u_i, v_j)^T$, where $0 \leq u_1 < u_2 < ... < u_{N_1} \leq 1$ and $0 \leq v_1 < v_2 < ... < v_{N_2} \leq 1$. Since it becomes possible to specify different types of exponential covariance structures in a multidimensional setting, Ying (1993) investigated two different cases. First, for a given separation vector $\mathbf{h} \equiv (h_1, h_2)^T$, they considered the stationary nuggetless covariance model $C_y(\mathbf{h}) = \sigma^2 \exp(-(\alpha_1|h_1| + \alpha_2|h_2|))$, and discovered that $\hat{\phi}_{\mathrm{ML}} = (\hat{\sigma}^2_{\mathrm{ML}}, \hat{\alpha}_{1,\mathrm{ML}}, \hat{\alpha}_{2,\mathrm{ML}})^T$ is asymptotically normally distributed with a $\sqrt{N_1}$ rate of convergence, provided

that $N_2/N_1 \to C \in (0,\infty)$. This is in stark contrast to the one-dimensional case; and is attributable to the separation of α_1 and α_2 by direction, allowing for asymptotic identifiability of the individual parameters. In contrast, under the nuggetless isotropic covariance model $C_y(\mathbf{h}) = \sigma^2 \exp(-\alpha\|\mathbf{h}\|)$ with $\|\cdot\|$ corresponding to Euclidean distance, σ^2 and α (still) cannot be distinguished asymptotically for a given $\alpha\sigma^2$.

2.3 Maximum Composite Likelihood Estimation in Geostatistics

The majority of composite likelihood functions that have been proposed in the literature are motivated by a search for more computationally efficient methods of estimating and performing inference on parameters in a geostatistical model. These functions are defined in such a way that the marginal or conditional densities that comprise the summand of the composite log-likelihood are relatively simple to obtain.

One of the earliest applications of a composite conditional likelihood is due to Besag (1974). Here, they considered taking the product of conditional densities, where the conditioning set contains only nearby observations.

Definition 2.1 (*Composite conditional K-nearest neighbours likelihood*) Let $\mathcal{B}_i$ contain the K observations $z(\mathbf{s}_j)$ where $\|\mathbf{s}_i - \mathbf{s}_j\|$ is smallest for $j \neq i$. Then $\mathcal{L}_C(\boldsymbol{\theta};\mathbf{z}) = \prod_{i=1}^{N} f(z(\mathbf{s}_i)|\mathcal{B}_i;\boldsymbol{\theta})$ is called a composite conditional K-nearest neighbours likelihood.

This composite likelihood was initially used by Besag (1974) in the context of a Markov process, where $(z(\mathbf{s}_i)|z(\mathbf{s}_1), z(\mathbf{s}_2), ..., z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i+1}), ..., z(\mathbf{s}_N)) \overset{d}{=} (z(\mathbf{s}_i)|\mathcal{B}_i)$ for some subset of nearby observations $\mathcal{B}_i$. As an example, they considered a spatial autoregressive model in two dimensions on an equally-spaced rectangular lattice, where $y(\mathbf{s}_i)$ is a linear combination of its immediate neighbours in the four cardinal directions, and an independent error term.

Another type of composite conditional likelihood is due to Vecchia (1988), where the conditioning set

$\mathcal{B}_i$ can only contain observations whose index is less than i; that is, the observations are ordered into a sequence, and $\mathcal{B}_i$ contains only a certain number of previous observations in the sequence. A specific choice of $\mathcal{B}_i$ is as defined below:

Definition 2.2 (*Composite conditional K-sequential neighbours likelihood*) For any ordering of the observation locations $\{\mathbf{s}_1,...,\mathbf{s}_N\}$, let the conditioning set $\mathcal{B}_1 = \emptyset$, $\mathcal{B}_i = \{z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i-2}),..., z(\mathbf{s}_1)\}$ for $2 \leq i \leq K$, and $\mathcal{B}_i = \{z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i-2}),..., z(\mathbf{s}_{i-K})\}$ for $i > K$. Then $\mathcal{L}_C(\boldsymbol{\theta};\mathbf{z}) = \prod_{i=1}^{N} f(z(\mathbf{s}_i)|\mathcal{B}_i;\boldsymbol{\theta}) = f(z(\mathbf{s}_1), z(\mathbf{s}_2),..., z(\mathbf{s}_K);\boldsymbol{\theta}) \prod_{i=K+1}^{N} f(z(\mathbf{s}_i)|z(\mathbf{s}_{i-1}), z(\mathbf{s}_{i-2}),..., z(\mathbf{s}_{i-K});\boldsymbol{\theta})$ is called a composite conditional K-sequential neighbours likelihood.

The motivation behind this composite likelihood is that it approximates the multiplicative law of probability, and in fact reconciles with the full likelihood if $K = N - 1$. An alternative choice of conditioning set that Vecchia (1988) considered is where $\mathcal{B}_i$ contains the K nearest observations to $\mathbf{s}_i$ from $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{i-1}\}$. This contrasts to Definition 2.1 which does not depend on the order of the observations.

Stein et al. (2004) extended the above work by investigating the effect of conditioning on observations that are not necessarily the nearest neighbours. This is done in the context of restricted maximum composite likelihood, where $\boldsymbol{\beta}$ is treated as a nuisance parameter vector and the focus is on the covariance parameters $\boldsymbol{\phi}$. They presented both theoretical and simulation-based examples comparing a nearest neighbours conditioning scheme to one that contains a mixture of nearby and distant observations. To do this, Stein et al. (2004) derived general expressions in a Gaussian geostatistical setting for $\mathbf{H}(\boldsymbol{\phi})$ and $\mathbf{J}(\boldsymbol{\phi})$ to form the sandwich covariance matrix $\mathbf{G}(\boldsymbol{\phi})^{-1}$. Comparison of these asymptotic variances showed that conditioning on some distant observations can lead to more statistically efficient parameter estimation. However, a remark made by both Vecchia (1988) and Stein et al. (2004) is that an optimal choice of conditioning sets is dependent on the true parameter values $\boldsymbol{\theta}_0$, so in practice there is little guidance on how to select such conditioning sets.

For composite marginal likelihood functions, Caragea and Smith (2007) and Oman and Landsman (2007) considered a setup where the observations are partitioned into B disjoint blocks.

Definition 2.3 (*Composite marginal blockwise likelihood*) Let $\mathcal{B}_1 \cup ... \cup \mathcal{B}_B = \{z(\mathbf{s}_1), z(\mathbf{s}_2), ..., z(\mathbf{s}_N)\}$, with $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for $i \neq j$. Then $\mathcal{L}_C(\boldsymbol{\theta}; \mathbf{z}) = \prod_{i=1}^{B} f(\mathcal{B}_i; \boldsymbol{\theta})$ is called a composite marginal blockwise likelihood.

This choice of composite likelihood involves misspecifying the dependence structure such that blocks of observations are independent of each other. Note that $B = 1$ corresponds to the full likelihood, and $B = N$ corresponds to assuming all of the observations are independent.

Caragea and Smith (2007) outlined an analytical approach to deriving expressions for $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ in a Gaussian spatial regression framework. This involves rewriting each mean-normalised $z(\mathbf{s}_i)$ in causal form (as is commonly seen in time series analysis); that is, $z(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\beta} = \sum_{r=0}^{\infty} c_r(\boldsymbol{\phi})\xi_{i-r}$, where $\xi_j \overset{iid}{\sim} N(0, v(\boldsymbol{\phi}))$. The Gaussian composite log-likelihood can then be written as a linear combination of $\xi_i\xi_j$ pairs, which simplifies the step of evaluating the expectations in $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$.

To illustrate this procedure, Caragea and Smith (2007) applied this to a one-dimensional Gaussian exponential covariance model in expanding domain, with B blocks of equal size partitioning the number line. For simplicity, they assumed that the data $\mathbf{z}$ had zero mean with σ^2 known, so that the only parameter of interest was α. Due to the equivalence of this situation with an autoregressive process of order one from time series (which will be demonstrated in Section 3.3), it is well-known that the causal form has coefficients given by $c_r(\phi) = \mathrm{e}^{-r\alpha}$, with $v(\phi) = \sigma^2(1 - \mathrm{e}^{-2\alpha})$ (see Cressie and Wikle (2011, p. 169) for instance). However, the calculation of $\mathbf{H}(\alpha)$ and $\mathbf{J}(\alpha)$ quickly becomes complicated as it involves four or more nested summations, so computation was used to perform this evaluation. It was shown that the efficiency of the maximum composite likelihood estimator relative to the maximum likelihood estimator for α in this situation is quite high, with performance improving as the number of blocks B decreases. This is due to the fact that smaller values of B more closely approximate the full likelihood. Oman and Landsman (2007) applied the composite marginal blockwise likelihood estimator in the context of binary spatial data and also found similarly strong performance in terms of relative efficiency.

Another class of composite likelihood function is the composite pairwise likelihood, which is composed of bivariate marginal or conditional densities. Since a bivariate vector is the simplest situation where dependence between observations can be introduced, the composite pairwise likelihood is a regularly chosen composite likelihood in situations where the data are not necessarily normally distributed (see Davis and Chun (2011) and Hui et al. (2018) for instance).

Definition 2.4 (*Composite pairwise likelihoods*) The function $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) = \prod_{i\neq j}^{B} f(z(\mathbf{s}_i)|z(\mathbf{s}_j);\boldsymbol{\theta})$ is called the composite conditional pairwise likelihood, and $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z}) = \prod_{i<j}^{B} f(z(\mathbf{s}_i),z(\mathbf{s}_j);\boldsymbol{\theta})$ is called the composite marginal pairwise likelihood.

Compared to the composite marginal blockwise likelihood with a block size of two, the composite marginal pairwise likelihood incorporates the joint density of all pairs of observations, and so there is no ambiguity in its construction.

In a geostatistical setting, the asymptotics of maximum composite pairwise likelihood estimation have been studied. Bevilacqua and Gaetan (2015) showed that under a Gaussian spatial regression setting, the maximum composite pairwise likelihood estimator is consistent and asymptotically normal in an expanding domain framework. Additionally, general expressions for $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ were derived. More recently, Bachoc et al. (2018) extended the work of Ying (1991) by exploring the infill asymptotics of maximum composite pairwise likelihood estimation under the one-dimensional nuggetless Gaussian exponential covariance model. They showed that, subject to some further conditions pertaining to the sampling scheme, $\hat{\alpha}_{\mathrm{CL}}\hat{\sigma}^2_{\mathrm{CL}}$ is consistent and asymptotically normal. Expressions for the asymptotic variance were also derived; though they involve four nested summations, and so comparisons to maximum likelihood estimation required computation.

The consistency and asymptotic normality of maximum composite likelihood estimation has also been studied and proven for other choices of composite likelihood. For instance, results can be found in Mardia, Hughes, et al. (2007) for the composite conditional K-nearest neighbours likelihood, and Oman and Landsman (2007) and Caragea and Smith (2007) for the composite marginal blockwise likelihood. In

essence, consistency and asymptotic normality often hold under similar regularity conditions to maximum likelihood estimation such as those in Theorems 2.1 and 2.3. In particular, a necessary condition for our usual asymptotics to hold is for the sandwich covariance matrix $\mathbf{G}(\boldsymbol{\theta})^{-1}$ to converge to the zero matrix (Varin, 2008).

However, the literature on maximum composite likelihood estimation in a geostatistical setting has limited results on the statistical performance of such estimators relative to maximum likelihood estimation, particularly from a theoretical standpoint. Earlier papers such as Besag (1974) and Vecchia (1988) focused on applying proposed composite likelihoods to estimate parameters in models for data from ecology, but have little discussion on the accuracy and precision of these estimates. It is only in more recent works such as Stein et al. (2004), Caragea and Smith (2007) and Bevilacqua and Gaetan (2015) that the asymptotic variance of maximum composite likelihood estimation has been investigated in greater detail by simplifying the form of the sandwich covariance matrix in a Gaussian geostatistical setting. Despite this, these analyses are mainly numerical in nature, which limits the feasibility of testing a wide variety of parameter values and model setups. Some closed-form expressions for the asymptotic relative efficiency of maximum composite likelihood estimation are available in Cox and Reid (2004) and Mardia, Hughes, et al. (2007). However, this is for Gaussian data where each pair of observations is assumed to have the same (unknown) correlation, which is often unrealistic.

To address this gap in the literature, we will derive some theoretical results on the asymptotic performance of maximum composite likelihood estimation relative to maximum likelihood estimation under the Gaussian exponential covariance model. This will be done by obtaining an exact expression for $\mathbf{G}(\boldsymbol{\theta})^{-1}$ for various choices of composite likelihood and analysing its convergence as the sample size N is taken to infinity. Relative asymptotic performance will be measured by taking the ratio of the entries of $\mathbf{I}(\boldsymbol{\theta})^{-1}$ to the entries of $\mathbf{G}(\boldsymbol{\theta})^{-1}$.

Chapter 3

Theory: Gaussian Exponential Covariance Model in One Dimension

The derivation of a closed-form expression for the sandwich covariance matrix of a maximum composite likelihood estimator is often tedious and requires complicated algebraic manipulation. Consequently, analysis of the statistical performance of maximum composite likelihood estimators has predominantly been performed numerically in the literature. However, this comes with the drawback of being unable to truly understand the key relationships and factors underpinning the statistical performance of these estimators.

In this chapter, we will derive and analyse the sandwich covariance matrix of maximum composite likelihood estimators in a one-dimensional zero-mean Gaussian nuggetless exponential covariance model with equally-spaced observations. In Section 3.1, we will introduce our lattice setup that will allow for analysis under all three frameworks in Definition 1.11. We will first consider maximum likelihood estimation and present a derivation of the inverse Fisher information matrix $\mathbf{I}(\phi)^{-1}$ in Section 3.2. As the main contribution of this thesis, we will then derive closed-form expressions for the sandwich covariance matrix $\mathbf{G}(\phi)^{-1}$ in two different cases: the composite conditional 2-nearest neighbours likelihood in Section 3.3 and the composite marginal blockwise likelihood in Section 3.4. By investigating their relative efficiencies, we will discover new insights that have not previously appeared in the literature.

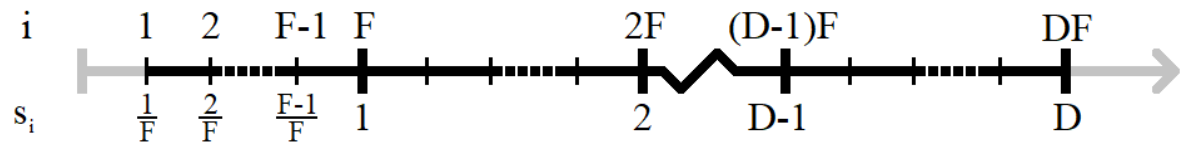

Figure 3.1: Setup of observation locations on a line for analysis under the hybrid asymptotic framework.

3.1 Equally-Spaced Lattice Construction

In order to unify the expanding domain and infill frameworks, which have often been treated separately in the literature, we will set up our equally-spaced lattice in a way that allows for analysis under the hybrid framework. To do so, we will consider an interval of length D, where each one-unit length is subdivided into F subintervals. This corresponds to the observation locations being defined as $s_i \equiv \frac{i}{F}$ for $i \in \{1,...,DF\}$, which is illustrated in Figure 3.1. Note that unlike Zhang and Zimmerman (2005), we will exclude the point $s_0 = 0$ in all of our derivations, but this will not affect the resulting asymptotic distribution (2.1). This exclusion is simply for convenience when working with the composite marginal blockwise likelihood in Section 3.4.

To prove that our setup is suitable under the various geostatistical asymptotic frameworks, we will show that they conform to Definition 1.11. From the notation used in Definition 1.11, the location furthest away from a given s_j is $\Delta_{j,N} = \max\left\{\frac{j-1}{F}, D - \frac{j}{F}\right\}$ units away, and the closest observation is $\delta_{j,N} = 1/F$ units away. Consequently,

$$\Delta_{DF} = \min_{1\le j\le DF} \max\left\{\frac{j-1}{F}, D - \frac{j}{F}\right\} \to \infty \qquad \text{as } D \to \infty,$$

and

$$\delta_{DF} = \max_{1\le j\le DF} \frac{1}{F} \to 0 \qquad \text{as } F \to \infty.$$

Thus, D controls the expanding domain aspect of the construction and F controls the frequency or density of observation locations to allow for infill analysis.

3.2 Full Likelihood

The Fisher information matrix for the one-dimensional nuggetless exponential covariance Gaussian process is presented in Zhang and Zimmerman (2005) under the expanding domain framework. What follows is an outline of the derivation of this matrix under the more general hybrid framework.

3.2.1 Derivation of the Inverse Fisher Information Matrix

Let $\mathbf{y} = (y(s_1), y(s_2), ..., y(s_{DF}))^T \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_{0,ij} = \text{cov}[y(s_i), y(s_j)] = \sigma_0^2 e^{-\alpha_0 |s_i - s_j|}$. The full log-likelihood can be written as

$$\ell(\sigma^2, \alpha; \mathbf{y}) = -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \text{constant}, \tag{3.1}$$

where

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & e^{-\frac{\alpha}{F}} & e^{-\frac{2\alpha}{F}} & \dots & e^{-\frac{(DF-2)\alpha}{F}} & e^{-\frac{(DF-1)\alpha}{F}} \\ e^{-\frac{\alpha}{F}} & 1 & e^{-\frac{\alpha}{F}} & \dots & e^{-\frac{(DF-3)\alpha}{F}} & e^{-\frac{(DF-2)\alpha}{F}} \\ e^{-\frac{2\alpha}{F}} & e^{-\frac{\alpha}{F}} & 1 & \dots & e^{-\frac{(DF-4)\alpha}{F}} & e^{-\frac{(DF-3)\alpha}{F}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-\frac{(DF-2)\alpha}{F}} & e^{-\frac{(DF-3)\alpha}{F}} & e^{-\frac{(DF-4)\alpha}{F}} & \dots & 1 & e^{-\frac{\alpha}{F}} \\ e^{-\frac{(DF-1)\alpha}{F}} & e^{-\frac{(DF-2)\alpha}{F}} & e^{-\frac{(DF-3)\alpha}{F}} & \dots & e^{-\frac{\alpha}{F}} & 1 \end{bmatrix}. \tag{3.2}$$

In order to find the determinant $|\mathbf{\Sigma}|$ and inverse $\mathbf{\Sigma}^{-1}$, Kac et al. (1953) derived the Cholesky decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{D}\mathbf{L}^T$, where $\mathbf{L}$ is unit lower triangular and $\mathbf{D}$ is diagonal. They showed that

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ e^{-\frac{\alpha}{F}} & 1 & 0 & \dots & 0 \\ e^{-\frac{2\alpha}{F}} & e^{-\frac{\alpha}{F}} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{-\frac{(DF-1)\alpha}{F}} & e^{-\frac{(DF-2)\alpha}{F}} & e^{-\frac{(DF-3)\alpha}{F}} & \dots & 1 \end{bmatrix},$$

and $\mathbf{D} = \sigma^2 \text{diag}\big(1, \underbrace{1-e^{-\frac{2\alpha}{F}}, 1-e^{-\frac{2\alpha}{F}}, \dots, 1-e^{-\frac{2\alpha}{F}}}_{DF-1 \text{ times}}\big)$. The determinant of $\mathbf{\Sigma}$ is then

$$|\mathbf{\Sigma}| = |\mathbf{L}||\mathbf{D}||\mathbf{L}^T| = |\mathbf{D}| = (\sigma^2)^{DF}(1-e^{-\frac{2\alpha}{F}})^{DF-1}. \tag{3.3}$$

To find the inverse of $\mathbf{\Sigma}$, we first invert $\mathbf{L}$ to obtain

$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -e^{-\frac{\alpha}{F}} & 1 & 0 & \dots & 0 & 0 \\ 0 & -e^{-\frac{\alpha}{F}} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 1 & 0 \\ 0 & 0 & 0 & \dots & -e^{-\frac{\alpha}{F}} & 1 \end{bmatrix},$$

so that

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{L}^{-1})^T \mathbf{D}^{-1} \mathbf{L}^{-1} = \frac{1}{\sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}})} \begin{bmatrix} 1 & -\mathrm{e}^{-\frac{\alpha}{F}} & 0 & \dots & 0 & 0 \\ -\mathrm{e}^{-\frac{\alpha}{F}} & 1+\mathrm{e}^{-\frac{2\alpha}{F}} & -\mathrm{e}^{-\frac{\alpha}{F}} & \dots & 0 & 0 \\ 0 & -\mathrm{e}^{-\frac{\alpha}{F}} & 1+\mathrm{e}^{-\frac{2\alpha}{F}} & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1+\mathrm{e}^{-\frac{2\alpha}{F}} & -\mathrm{e}^{-\frac{\alpha}{F}} \\ 0 & 0 & 0 & \dots & -\mathrm{e}^{-\frac{\alpha}{F}} & 1 \end{bmatrix},$$

or in terms of the entries,

$$\{\boldsymbol{\Sigma}^{-1}\}_{ij} = \frac{1}{\sigma^2\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)} \begin{cases} 1, & i=j \in \{1, DF\} \\ 1+\mathrm{e}^{-\frac{2\alpha}{F}}, & i=j \in \{2,3,...,DF-1\} \\ -\mathrm{e}^{-\frac{\alpha}{F}}, & |i-j|=1 \\ 0, & \text{otherwise} \end{cases}.$$

Using (3.3), we can then rewrite (3.1) as

$$\ell(\sigma^2, \alpha; \mathbf{y}) = -\frac{DF}{2}\log(\sigma^2) - \frac{DF-1}{2}\log(1-\mathrm{e}^{-\frac{2\alpha}{F}}) - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \text{constant},$$

so the score function is given by

$$\mathrm{sc}(\sigma^2, \alpha; \mathbf{y}) = \begin{bmatrix} \frac{\partial \ell}{\partial \sigma^2} \\ \frac{\partial \ell}{\partial \alpha} \end{bmatrix} = \begin{bmatrix} -\frac{DF}{2\sigma^2} - \frac{1}{2}\mathbf{y}^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \sigma^2}\mathbf{y} \\ -\frac{(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}}{F\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)} - \frac{1}{2}\mathbf{y}^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \alpha}\mathbf{y} \end{bmatrix}.$$

Note that the system of equations $\mathrm{sc}(\sigma^2, \alpha; \mathbf{y}) = \mathbf{0}$ cannot be solved analytically, so the maximum likelihood estimates $\hat{\sigma}^2$ and $\hat{\alpha}$ need to be found numerically.

For asymptotic analysis of these estimators, we are interested in computing the Fisher information ma-

trix. Firstly, we require the (observed) information matrix:

$$\text{info}(\sigma^2,\alpha;\mathbf{y}) = -\begin{bmatrix} \frac{\partial^2\ell}{\partial(\sigma^2)^2} & \frac{\partial^2\ell}{\partial\sigma^2\partial\alpha} \\ \frac{\partial^2\ell}{\partial\sigma^2\partial\alpha} & \frac{\partial^2\ell}{\partial\alpha^2} \end{bmatrix} = \begin{bmatrix} -\frac{DF}{2\sigma^4} + \frac{1}{2}\mathbf{y}^T\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial(\sigma^2)^2}\mathbf{y} & \frac{1}{2}\mathbf{y}^T\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\sigma^2\partial\alpha}\mathbf{y} \\ \frac{1}{2}\mathbf{y}^T\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\sigma^2\partial\alpha}\mathbf{y} & \frac{2(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}}{F^2\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)^2} + \frac{1}{2}\mathbf{y}^T\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\alpha^2}\mathbf{y} \end{bmatrix}.$$

In order to compute the expectations of these terms, we will make use of the following lemma:

Lemma 3.1 (*Expectation of a quadratic form*) Let $\mathbf{x} = (x_1,...,x_n)^T$ follow a joint distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Also, let $\mathbf{U} = (u_{ij})_{n\times n}$. Then $\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{x}] = \text{tr}(\mathbf{U}\boldsymbol{\Sigma})$, where $\text{tr}(\cdot)$ is the trace operator.

Proof: $\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{x}] = \sum_{i=1}^n\sum_{j=1}^n u_{ij}\mathbb{E}[x_ix_j] = \sum_{i=1}^n\sum_{j=1}^n u_{ij}\Sigma_{ji} = \sum_{i=1}^n(\mathbf{U}\boldsymbol{\Sigma})_{ii} = \text{tr}(\mathbf{U}\boldsymbol{\Sigma})$.

By applying this lemma to the top-left element of the information matrix, we obtain

$$\begin{aligned}\mathbb{E}\left[-\frac{DF}{2\sigma^4} + \frac{1}{2}\mathbf{y}^T\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial(\sigma^2)^2}\mathbf{y}\right] &= -\frac{DF}{2\sigma^4} + \frac{1}{2}\text{tr}\left(\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial(\sigma^2)^2}\boldsymbol{\Sigma}\right) \\ &= -\frac{DF}{2\sigma^4} + \frac{1}{2}\text{tr}\left(\frac{2}{\sigma^4}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\ &= -\frac{DF}{2\sigma^4} + \frac{DF}{\sigma^4} = \frac{DF}{2\sigma^4}.\end{aligned}$$

In this example, computation of the second-order partial derivative with respect to σ^2 was straightforward since it only appears as a multiplicative term. Calculation of the other terms follows in a similar manner, but requires slightly more involved algebraic manipulation. Ultimately, one can derive the entire (expected) Fisher information matrix, which is

$$\mathbf{I}(\sigma^2,\alpha;\mathbf{y}) = \mathbb{E}[\text{info}(\sigma^2,\alpha)] = \begin{bmatrix} \frac{DF}{2(\sigma^2)^2} & \frac{(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}}{\sigma^2F\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)} \\ \frac{(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}}{\sigma^2F\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)} & \frac{(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}\left(1+\mathrm{e}^{-\frac{2\alpha}{F}}\right)}{F^2\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)^2} \end{bmatrix}. \tag{3.4}$$

This conforms with the Fisher information matrix presented in Zhang and Zimmerman (2005), albeit

with DF observations instead of $DF+1$. The inverse of the Fisher information matrix is then given by

$$\mathbf{I}(\sigma^2,\alpha)^{-1} = \frac{1}{\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)DF + 2\mathrm{e}^{-\frac{2\alpha}{F}}} \begin{bmatrix} 2(\sigma^2)^2\left(1+\mathrm{e}^{-\frac{2\alpha}{F}}\right) & -2\sigma^2F\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right) \\ -2\sigma^2F\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right) & F^2\frac{DF\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)^2}{(DF-1)\mathrm{e}^{-\frac{2\alpha}{F}}} \end{bmatrix}. \tag{3.5}$$

3.2.2 Asymptotics

Under expanding domain asymptotics ($D\to\infty$ and F fixed), each of the terms in (3.5) converges to 0 at a rate of $1/D$, and so the maximum likelihood estimators of σ^2 and α are consistent. Furthermore, the attainment of asymptotic normality can be verified by applying Theorem 2.2. Specifically, we have the set of all unique displacements $\mathcal{H}_N \equiv \{|s_i - s_j|; i,j \in \{1,2,...,N\}\} = \{0, \frac{1}{F}, \frac{2}{F}, ..., \frac{DF-1}{F}\}$ and covariance function $C_y(h) = \sigma^2\mathrm{e}^{-\alpha|h|}$, so $\lim_{D\to\infty}\sum_{\mathbf{h}\in\mathcal{H}_N}|C_y(h)| = \lim_{D\to\infty}\sum_{j=0}^{DF-1}\sigma^2\mathrm{e}^{-\frac{\alpha j}{F}}$ is a convergent geometric sum, and the first and second-order partial derivatives of $C_y(\mathbf{h})$ also form convergent sums or equal zero. Thus, by noting that

$$\lim_{D\to\infty} DF\mathbf{I}(\sigma^2,\alpha)^{-1} = \begin{bmatrix} 2(\sigma^2)^2\frac{1+\mathrm{e}^{-\frac{2\alpha}{F}}}{1-\mathrm{e}^{-\frac{2\alpha}{F}}} & -2F\sigma^2 \\ -2F\sigma^2 & F^2(\mathrm{e}^{\frac{2\alpha}{F}}-1) \end{bmatrix} \equiv \mathbf{V}(\sigma^2,\alpha), \tag{3.6}$$

an asymptotic distribution for large values of D is

$$\sqrt{DF}\begin{bmatrix} \hat{\sigma}^2_{\mathrm{ML}} - \sigma^2_0 \\ \hat{\alpha}_{\mathrm{ML}} - \alpha_0 \end{bmatrix} \dot{\sim} N(\mathbf{0}, \mathbf{V}(\sigma^2_0, \alpha_0)).$$

This matches up with the work of Zhang and Zimmerman (2005) and (2.1).

In contrast, under infill asymptotics ($F\to\infty$ and D fixed), both $\hat{\sigma}^2_{\mathrm{ML}}$ and $\hat{\alpha}_{\mathrm{ML}}$ do not conform to the usual asymptotic behaviour of maximum likelihood estimators. This can be shown by taking the limit of the terms in (3.5) with respect to F and using the Taylor expansion $\mathrm{e}^{ah} = 1 + ah + a^2h^2/2 + O(h^3)$.

As an example, the limit of the top-left element of the inverse Fisher information matrix can be found as follows:

$$\begin{aligned}
&\lim_{F\to\infty} \frac{2(\sigma^2)^2\left(1+\mathrm{e}^{-\frac{2\alpha}{F}}\right)}{\left(1-\mathrm{e}^{-\frac{2\alpha}{F}}\right)DF+2\mathrm{e}^{-\frac{2\alpha}{F}}} = \lim_{\delta\to 0} \frac{2\delta(\sigma^2)^2(1+\mathrm{e}^{-2\alpha\delta})}{(1-\mathrm{e}^{-2\alpha\delta})D+2\delta\mathrm{e}^{-2\alpha\delta}} \\
&= \lim_{\delta\to 0} \frac{2\delta(\sigma^2)^2(1+(1-2\alpha\delta+2\alpha^2\delta^2+O(\delta^3)))}{(1-(1-2\alpha\delta+2\alpha^2\delta^2+O(\delta^3)))D+2\delta(1-2\alpha\delta+2\alpha^2\delta^2+O(\delta^3))} \\
&= \lim_{\delta\to 0} \frac{2(\sigma^2)^2(\delta-\alpha\delta^2+\alpha^2\delta^3+O(\delta^4))}{(\alpha\delta-\alpha^2\delta^2+O(\delta^3))D+\delta(1-2\alpha\delta+2\alpha^2\delta^2+O(\delta^3))} \\
&= \lim_{\delta\to 0} \frac{2(\sigma^2)^2(1+O(\delta))}{(\alpha+O(\delta))D+1+O(\delta)} = \frac{2(\sigma^2)^2}{\alpha D+1}.
\end{aligned}$$

By applying the same procedure to each of the terms in (3.5), we obtain

$$\lim_{F\to\infty} \mathbf{I}(\sigma^2,\alpha)^{-1} = \frac{2}{\alpha D+1}\begin{bmatrix} (\sigma^2)^2 & -\alpha\sigma^2 \\ -\alpha\sigma^2 & \alpha^2 \end{bmatrix}. \tag{3.7}$$

Here, none of the elements converge to zero, so the consistency of $\hat{\sigma}^2_{\mathrm{ML}}$ and $\hat{\alpha}_{\mathrm{ML}}$ is not guaranteed (Ying, 1991). Some insight into why this is the case can be gained by considering the interaction between the variability of estimation in $\hat{\sigma}^2_{\mathrm{ML}}$ and the true value of α. In both (3.6) and (3.7), which correspond to the limit of the inverse Fisher information matrix under the expanding domain and infill approaches respectively, the top-left element is a monotone decreasing function of α. This suggests that if there is a particularly strong correlation between nearby observations ($\alpha_0 \approx 0$), then $\hat{\sigma}^2_{\mathrm{ML}}$ will be subject to a higher level of variability than if the observations have a low correlation (α_0 large). Hence, in the case of infill asymptotics where observations are taken near each other and are thus strongly dependent, each additional observation that is sampled becomes far less informative about the true value of σ^2.

Another factor which contributes to the behaviour of infill asymptotics in this case is the fact that both σ^2 and α are being estimated simultaneously. If we assume that α is known, then we need only consider the top-left element in (3.4) to find the asymptotic variance of $\hat{\sigma}^2_{\mathrm{ML}}$, which is $\mathrm{var}(\hat{\sigma}^2_{\mathrm{ML}}) \approx 2(\sigma^2_0)^2/(DF)$. Clearly, this converges to zero under both the expanding domain and infill frameworks. Similarly, if

we assume that σ^2 is known, then we need only consider the bottom-right element in (3.4) to find the asymptotic variance of $\hat{\alpha}_{\mathrm{ML}}$, which also converges to zero under both asymptotic frameworks. Based on this, the asymptotic distributions for $\hat{\sigma}^2_{\mathrm{ML}}$ and $\hat{\alpha}_{\mathrm{ML}}$ as shown by Ying (1991) can be derived, which results in (2.2) and (2.3). However, we know from (3.7) that convergence of the variance does not occur under infill asymptotics when both parameters are unknown. This can be attributed to the fact that having to estimate both parameters simultaneously effectively leads to them competing for information in the data.

In order for the usual asymptotic results to hold while using an infill sampling scheme, it will be necessary to consider the hybrid framework where the domain of sampling will expand at the same time. From (3.7), it is clear that as long as $D \to \infty$, irrespective of the rate, the inverse Fisher information matrix will approach the zero matrix. This suggests that, when the domain is of a finite size, there is asymptotically insufficient information about σ^2 and α to achieve consistent estimation simultaneously.

3.3 Composite Conditional 2-Nearest Neighbours Likelihood

For the classes of composite conditional likelihood that we have highlighted in Section 2.3, we can first consider the composite conditional K-sequential neighbours likelihood. It is well-known (see Cressie and Wikle (2011, p. 169) for instance) that the one-dimensional exponential covariance Gaussian process on an equally-spaced lattice can be equivalently expressed as an autoregressive process of order 1 from time series, where $y(s_1) \sim N(0, \sigma^2)$ and

$$(y(s_i)|y(s_1),...,y(s_{i-1})) \overset{d}{=} (y(s_i)|y(s_{i-1})) \sim N(\mathrm{e}^{-\frac{\alpha}{F}} y(s_{i-1}), \sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}})), \qquad 1 < i \leq N = DF.$$

Since this process satisfies the Markov property, the constructed composite likelihood will be equivalent for all $K \geq 1$, and is in turn equivalent to the full likelihood. However, it is important to note that this equivalence to the full likelihood is only true under the "natural" ordering of observations that

is available on a number line. Thus, a potential issue with the composite conditional K-sequential neighbours likelihood is the choice of observation sequence, which is more prominent in two or higher dimensions. In light of this, we shall investigate the statistical performance of the composite conditional K-nearest neighbours likelihood as its construction is far less ambiguous.

3.3.1 Construction of the Composite Likelihood

In order to find conditional distributions of Gaussian random variables, we can apply the following lemma, as presented for example by Eaton (1983, p. 116-117):

Lemma 3.2 (*Conditional Gaussian distribution*) Let $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, and $\mathbf{x} \equiv (\mathbf{x}_1^T, \mathbf{x}_2^T)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then $(\mathbf{x}_1|\mathbf{x}_2) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Now under the one-dimensional exponential covariance Gaussian process on an equally-spaced lattice, it can shown using this lemma that

$$(y(s_1)|y(s_2), ..., y(s_{K+1})) \stackrel{d}{=} (y(s_1)|y(s_2)) \sim N(\mathrm{e}^{-\frac{\alpha}{F}} y(s_2), \sigma^2(1 - \mathrm{e}^{-\frac{2\alpha}{F}})), \tag{3.8}$$

$$(y(s_{DF})|y(s_{DF-1}), ..., y(s_{DF-K})) \stackrel{d}{=} (y(s_{DF})|y(s_{DF-1})) \sim N(\mathrm{e}^{-\frac{\alpha}{F}} y(s_{DF-1}), \sigma^2(1 - \mathrm{e}^{-\frac{2\alpha}{F}})), \tag{3.9}$$

and for $K \geq 2$ and $1 < i \leq N - 1 = DF - 1$ that

$$(y(s_i)|K\text{-nearest neighbours}) \stackrel{d}{=} (y(s_i)|y(s_{i-1}), y(s_{i+1})) \sim N\left(\frac{\mathrm{e}^{-\frac{\alpha}{F}}(y(s_{i-1}) + y(s_{i+1}))}{1 + \mathrm{e}^{-\frac{2\alpha}{F}}}, \sigma^2 \frac{1 - \mathrm{e}^{-\frac{2\alpha}{F}}}{1 + \mathrm{e}^{-\frac{2\alpha}{F}}}\right). \tag{3.10}$$

This is a well-known result in the geostatistical literature; see Cressie and Wikle (2011, p. 170-171) for

instance. Thus, in the construction of the composite conditional K-nearest neighbours likelihood, all choices of $K \geq 2$ lead to the same composite likelihood. As such, it is adequate to simply analyse the cases where $K = 1$ and $K = 2$.

When $K = 1$, note that due to the equally-spaced lattice setup, all observation locations except for the endpoints s_1 and s_n have a non-unique nearest neighbour. In a real set of geostatistical data, this would rarely occur as observation locations are usually unequally spaced. However, if necessary, a method of dealing with this could involve the random selection of an observation in the case of a tie. Nevertheless, in our particular problem, we could break a tie by choosing to condition on the point to the left on the number line, leading to the composite likelihood $\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{y}) = f(y(s_1)|y(s_2);\boldsymbol{\theta})\prod_{i=2}^{N} f(y(s_i)|y(s_{i-1});\boldsymbol{\theta})$, which only differs from the full likelihood by the first term $f(y(s_1)|y(s_2);\boldsymbol{\theta})$. We would therefore expect the performance of maximum composite likelihood estimation based on this composite likelihood to quickly match the performance of maximum likelihood estimation as N increases and the effect of the slight perturbation diminishes.

The case of $K = 2$ is far more complicated and will be the focus of the remainder of this section. To begin, by using (3.8), (3.9) and (3.10), the composite conditional 2-nearest neighbours likelihood can be expressed as

$$\begin{aligned}
\mathcal{L}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{y}) &= f(y(s_1)|y(s_2);\boldsymbol{\theta}) f(y(s_{DF})|y(s_{DF-1});\boldsymbol{\theta}) \prod_{i=2}^{DF-1} f(y(s_i)|y(s_{i-1}),y(s_{i+1});\boldsymbol{\theta}) \\
&= \frac{1}{(2\pi\sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}}))^{\frac{DF}{2}}} \exp\Big(-\frac{(y(s_1)-\mathrm{e}^{-\frac{\alpha}{F}}y(s_2))^2+(y(s_{DF})-\mathrm{e}^{-\frac{\alpha}{F}}y(s_{DF-1}))^2}{2\sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}})}\Big) \\
&\quad \times \prod_{i=2}^{DF-1} \sqrt{1+\mathrm{e}^{-\frac{2\alpha}{F}}} \exp\Big(-\frac{1+\mathrm{e}^{-\frac{2\alpha}{F}}}{2\sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}})}\big(y(s_i)-\frac{\mathrm{e}^{-\frac{\alpha}{F}}}{1+\mathrm{e}^{-\frac{2\alpha}{F}}}(y(s_{i-1})+y(s_{i+1}))\big)^2\Big) \\
&= \frac{(1+\mathrm{e}^{-\frac{2\alpha}{F}})^{\frac{DF}{2}-1}}{(2\pi\sigma^2(1-\mathrm{e}^{-\frac{2\alpha}{F}}))^{\frac{DF}{2}}} \exp\Big(-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}\mathbf{y}\Big), \qquad (3.11)
\end{aligned}$$

where $\mathbf{y} = (y(s_1), ..., y(s_{DF}))^T$, and $\mathbf{M} \in \mathbb{R}^{DF \times DF}$ has the following pentadiagonal and symmetric form:

$$\mathbf{M} = \begin{bmatrix} a & d & e & & & & & & \\ d & b & d & e & & & & & \\ e & d & c & d & e & & & & \\ & e & d & c & d & \ddots & & & \\ & & e & d & c & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & c & d & e \\ & & & & & \ddots & d & b & d \\ & & & & & & e & d & a \end{bmatrix}.$$

The form of $\mathbf{M}$, which we shall call the composition matrix, means that we can decompose it into a linear combination of the following five simple matrices:

$$\mathbf{I} = \begin{bmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & & & & & & & \\ & 0 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & & & & & 0 & \\ & & & & & & & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & & & & & & & \\ 1 & 0 & 1 & & & & & & \\ & 1 & 0 & 1 & & & & & \\ & & 1 & 0 & \ddots & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & \ddots & 0 & 1 & & \\ & & & & & 1 & 0 & 1 & \\ & & & & & & 1 & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 0 & 0 & 1 & & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ 1 & 0 & 0 & 0 & \ddots & & & & \\ & 1 & 0 & 0 & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & 0 & 0 & 1 & \\ & & & & \ddots & 0 & 0 & 0 & \\ & & & & & 1 & 0 & 0 & \end{bmatrix}. \tag{3.12}$$

Hence, we can write

$$\mathbf{M} = \frac{1}{1-\rho^2}\left[\left(1+\frac{\rho^2}{1+\rho^2}\right)\mathbf{I} + 2\rho^2\mathbf{A} + \left(-\rho^2+\frac{\rho^2}{1+\rho^2}\right)\mathbf{B} - 2\rho\mathbf{C} + \frac{\rho^2}{1+\rho^2}\mathbf{D}\right], \tag{3.13}$$

where $\rho = e^{-\frac{\alpha}{F}}$.

3.3.2 Derivation of the Sandwich Covariance Matrix

Using (3.11), the composite conditional 2-nearest neighbours log-likelihood is given by

$$c\ell(\sigma^2,\alpha;\mathbf{y}) = -\frac{DF}{2}\log\sigma^2 - \frac{DF}{2}\log\left(1-e^{-\frac{2\alpha}{F}}\right) + \frac{DF-2}{2}\log\left(1+e^{-\frac{2\alpha}{F}}\right) - \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}\mathbf{y} + \text{const.}.$$

Noting that $\mathbf{M}$ is a function of α, this leads to the composite score function

$$\text{sc}_C(\sigma^2,\alpha;\mathbf{y}) = \begin{bmatrix} -\frac{DF}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}\mathbf{y} \\ -\frac{\rho^2}{F}\left(\frac{DF}{1-\rho^2} + \frac{DF-2}{1+\rho^2}\right) - \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} \end{bmatrix} \tag{3.14}$$

and the composite information function

$$\text{info}_C(\sigma^2,\alpha;\mathbf{y}) = \begin{bmatrix} -\frac{DF}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\mathbf{y}^T\mathbf{M}\mathbf{y} & -\frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} \\ -\frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} & -\frac{2\rho^2}{F^2}\left(\frac{DF}{(1-\rho^2)^2} + \frac{DF-2}{(1+\rho^2)^2}\right) + \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}''\mathbf{y} \end{bmatrix}, \tag{3.15}$$

where the first and second derivatives of $\mathbf{M}$ with respect to α are

$$\mathbf{M}' = -\frac{2\rho}{F(1-\rho^2)^2}\left[\left(\rho + \frac{\rho(1+\rho^4)}{(1+\rho^2)^2}\right)\mathbf{I} + 2\rho\mathbf{A} + \left(-\rho + \frac{\rho(1+\rho^4)}{(1+\rho^2)^2}\right)\mathbf{B} - (1+\rho^2)\mathbf{C} + \frac{\rho(1+\rho^4)}{(1+\rho^2)^2}\mathbf{D}\right]$$

and

$$\begin{aligned}\mathbf{M}'' = \frac{4\rho}{F^2(1-\rho^2)^3}\Bigg[&\left(\rho(1+\rho^2) + \frac{\rho(1+6\rho^4+\rho^8)}{(1+\rho^2)^3}\right)\mathbf{I} + 2\rho(1+\rho^2)\mathbf{A} \\ &+ \left(-\rho(1+\rho^2) + \frac{\rho(1+6\rho^4+\rho^8)}{(1+\rho^2)^3}\right)\mathbf{B} - \frac{1}{2}(1+6\rho^2+\rho^4)\mathbf{C} + \frac{\rho(1+6\rho^4+\rho^8)}{(1+\rho^2)^3}\mathbf{D}\Bigg].\end{aligned}$$

The objective now is to derive expressions for the two components of the sandwich variance: $\mathbf{H}(\sigma^2,\alpha) = \mathbb{E}[\text{info}_C(\sigma^2,\alpha;\mathbf{y})]$ and $\mathbf{J}(\sigma^2,\alpha) = \mathbb{E}[\text{sc}_C(\sigma^2,\alpha;\mathbf{y})\text{sc}_C(\sigma^2,\alpha;\mathbf{y})^T]$. First, to find $\mathbf{H}(\sigma^2,\alpha)$, we note the following trace formulae for the exponential covariance matrix $\mathbf{\Sigma}$ whose form is shown in (3.2):

$$\text{tr}(\mathbf{I\Sigma}) = \sigma^2 DF, \qquad \text{tr}(\mathbf{A\Sigma}) = \sigma^2(DF-2), \qquad \text{tr}(\mathbf{B\Sigma}) = \sigma^2(DF-4),$$
$$\text{tr}(\mathbf{C\Sigma}) = 2\sigma^2\rho(DF-1), \quad \text{tr}(\mathbf{D\Sigma}) = 2\sigma^2\rho^2(DF-2). \tag{3.16}$$

By Lemma 3.1 and the linearity of the trace function, we find, for instance by using (3.13), that

$$\begin{aligned}\mathbb{E}[\mathbf{y}^T\mathbf{M}\mathbf{y}] &= \text{tr}(\mathbf{M\Sigma}) \\ &= \frac{1}{1-\rho^2}\Bigg[\Big(1+\frac{\rho^2}{1+\rho^2}\Big)\text{tr}(\mathbf{I\Sigma}) + 2\rho^2\text{tr}(\mathbf{A\Sigma}) \\ &\qquad + \Big(-\rho^2+\frac{\rho^2}{1+\rho^2}\Big)\text{tr}(\mathbf{B\Sigma}) - 2\rho\text{tr}(\mathbf{C\Sigma}) + \frac{\rho^2}{1+\rho^2}\text{tr}(\mathbf{D\Sigma})\Bigg] \\ &= \sigma^2 DF. \end{aligned} \tag{3.17}$$

Similarly, it can be shown that

$$\mathbb{E}[\mathbf{y}^T\mathbf{M}'\mathbf{y}] = -\sigma^2\frac{4}{F(1-\rho^4)}\rho^2[DF-(1-\rho^2)]$$

and

$$\mathbb{E}[\mathbf{y}^T\mathbf{M}''\mathbf{y}] = \sigma^2\frac{4}{F^2(1-\rho^4)^2}\rho^2[(3+\rho^2+5\rho^4-\rho^6)DF-(1-\rho^2)(3-2\rho^2+3\rho^4)].$$

Thus, using (3.15), we obtain

$$\mathbf{H}(\sigma^2,\alpha) = \begin{bmatrix} \frac{DF}{2(\sigma^2)^2} & \frac{2\rho^2}{F\sigma^2(1-\rho^4)}[DF-(1-\rho^2)] \\ \frac{2\rho^2}{F\sigma^2(1-\rho^4)}[DF-(1-\rho^2)] & \frac{2\rho^2}{F^2(1-\rho^4)^2}[(1+\rho^2+3\rho^4-\rho^6)DF-(1-\rho^2)(1+3\rho^4)] \end{bmatrix}. \tag{3.18}$$

The calculation of $\mathbf{J}(\sigma^2,\alpha) = \mathbb{E}[\text{sc}_\text{C}(\sigma^2,\alpha;\mathbf{y})\text{sc}_\text{C}(\sigma^2,\alpha;\mathbf{y})^T]$ is considerably more complicated as it requires finding expressions for fourth-order moments. To begin, we consider the following lemma, as presented for example by Rencher and Schaalje (2008, p. 109):

Lemma 3.3 (*Covariance of a Gaussian quadratic form*) Let $\mathbf{x} = (x_1,...,x_n)^T$ follow a joint Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}$. Also, let $\mathbf{U}$ and $\mathbf{V}$ be $n \times n$ symmetric matrices.

Then $\text{cov}(\mathbf{x}^T\mathbf{U}\mathbf{x}, \mathbf{x}^T\mathbf{V}\mathbf{x}) = 2\text{tr}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Sigma})$.

Note the additional restrictions that have been imposed compared to Lemma 3.1; in particular the requirement for $\mathbf{x}$ to be Gaussian. This highlights a wider issue in the literature that results pertaining to asymptotic properties of maximum composite likelihood estimation under non-normal models are scarce: expressions for fourth-order moments are often complicated.

By applying both Lemma 3.1 and 3.3, it is straightforward to show the following useful result:

Corollary 3.4 (*Expectation of a Gaussian quartic form*) Let $\mathbf{x} = (x_1, ..., x_n)^T$ satisfy the conditions outlined in Lemma 3.3. Then $\mathbb{E}[\mathbf{x}^T\mathbf{U}\mathbf{x}\mathbf{x}^T\mathbf{V}\mathbf{x}] = \text{tr}(\mathbf{U}\boldsymbol{\Sigma})\text{tr}(\mathbf{V}\boldsymbol{\Sigma}) + 2\text{tr}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Sigma})$.

Hence, the task now shifts to finding matrix traces of the form $\text{tr}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Sigma})$. Once again due to linearity, we can first find traces of this form in terms of the five simple matrices in (3.12). In fact, we only need to find $\binom{5}{1} + \binom{5}{2} = 15$ of these instead of $5^2 = 25$ due to the cyclical invariance of traces; that is, $\text{tr}(\mathbf{UVWX}) = \text{tr}(\mathbf{XUVW})$ for matrices of conformable dimensions.

To simplify notation, define $u_n \equiv \sum_{k=1}^{n}(n-k)\rho^{2k}$. Also, let "$\circ$" denote the Hadamard (entrywise) product of two matrices. Then based on the exponential covariance matrix $\boldsymbol{\Sigma}$, we have as an example that

$$\begin{aligned}
\text{tr}(\mathbf{I}\boldsymbol{\Sigma}\mathbf{I}\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}) &= \sum_{i=1}^{DF}\sum_{j=1}^{DF}\{\boldsymbol{\Sigma}\circ\boldsymbol{\Sigma}^T\}_{ij} = \sum_{i=1}^{DF}\sum_{j=1}^{DF}(\sigma^2\rho^{|i-j|})^2 \\
&= \sigma^4\left[\sum_{i=1}^{DF}1 + \sum_{i\neq j}^{DF}\rho^{2|i-j|}\right] = \sigma^4\left[DF + 2\sum_{j=1}^{DF-1}\sum_{i=j+1}^{DF}\rho^{2|i-j|}\right] \\
&= \sigma^4\left[DF + 2\sum_{j=1}^{DF-1}\sum_{k=1}^{DF-j}\rho^{2k}\right] = \sigma^4\left[DF + 2\sum_{k=1}^{DF-1}\sum_{j=1}^{DF-k}\rho^{2k}\right] \\
&= \sigma^4\left[DF + 2\sum_{k=1}^{DF}(DF-k)\rho^{2k}\right] = \sigma^4[DF + 2u_{DF}].
\end{aligned}$$

Similarly, we can obtain all of the following:

$$\begin{aligned}
&\mathrm{tr}(\mathbf{I\Sigma I\Sigma}) = \sigma^4[DF + 2u_{DF}], && \mathrm{tr}(\mathbf{I\Sigma A\Sigma}) = \sigma^4[DF - 2 + 2u_{DF-1}],\\
&\mathrm{tr}(\mathbf{I\Sigma B\Sigma}) = \sigma^4[(1+2\rho^2)(DF-4) + 2\rho^2 u_{DF-3}], && \mathrm{tr}(\mathbf{I\Sigma C\Sigma}) = \sigma^4[4\rho(DF-1) + 4\rho u_{DF-1}],\\
&\mathrm{tr}(\mathbf{I\Sigma D\Sigma}) = \sigma^4[2\rho^2(DF-2) + 4u_{DF-1}], && \mathrm{tr}(\mathbf{A\Sigma A\Sigma}) = \sigma^4[DF - 2 + 2u_{DF-2}],\\
&\mathrm{tr}(\mathbf{A\Sigma B\Sigma}) = \sigma^4[DF - 4 + 2u_{DF-3}], && \mathrm{tr}(\mathbf{A\Sigma C\Sigma}) = \sigma^4[4\rho(DF-2) + 4\rho u_{DF-2}],\\
&\mathrm{tr}(\mathbf{A\Sigma D\Sigma}) = \sigma^4[2\rho^2(DF-2) + 4u_{DF-2}], && \mathrm{tr}(\mathbf{B\Sigma B\Sigma}) = \sigma^4[DF - 4 + 2u_{DF-4}],\\
&\mathrm{tr}(\mathbf{B\Sigma C\Sigma}) = \sigma^4[4\rho(DF-4) + 4\rho u_{DF-3}], && \mathrm{tr}(\mathbf{B\Sigma D\Sigma}) = \sigma^4[2\rho^2(DF-4) + 4u_{DF-3}],\\
&\mathrm{tr}(\mathbf{C\Sigma C\Sigma}) = \sigma^4[2(1+\rho^2)(DF-1) + 8u_{DF-1}], && \mathrm{tr}(\mathbf{C\Sigma D\Sigma}) = \sigma^4[4\rho(1+\rho^2)(DF-2) + 8\rho u_{DF-2}],\\
&\mathrm{tr}(\mathbf{D\Sigma D\Sigma}) = \sigma^4[2(1+\rho^4)(DF-2) + 4\rho^2(1+\rho^2)(DF-3) + 8\rho^2 u_{DF-3}]. && (3.19)
\end{aligned}$$

Using the above results, we can then find $\mathrm{tr}(\mathbf{M\Sigma M\Sigma})$, $\mathrm{tr}(\mathbf{M\Sigma M'\Sigma})$ and $\mathrm{tr}(\mathbf{M'\Sigma M'\Sigma})$. For instance by using (3.13), we have

$$\begin{aligned}
\mathrm{tr}(\mathbf{M\Sigma M\Sigma}) &= \frac{1}{(1-\rho^2)^2}\left[\left(1+\frac{\rho^2}{1+\rho^2}\right)^2 \mathrm{tr}(\mathbf{I\Sigma I\Sigma}) + 2\left(1+\frac{\rho^2}{1+\rho^2}\right)\times 2\rho^2 \mathrm{tr}(\mathbf{I\Sigma A\Sigma}) + ...\right]\\
&= \frac{\sigma^4}{(1+\rho^2)^2}[(1+4\rho^2+\rho^4)DF - 2\rho^2 + 4\rho^4]. \qquad (3.20)
\end{aligned}$$

The algebra required has been placed in the Appendix as it is lengthy; though a useful recursive relation which simplifies the calculation of (3.20) is that $u_{n+1} = \rho^2(u_n + n)$. In fact, observe that (3.20) is a linear function of D, which indicates that all of the non-linear terms u_n that are present in (3.19) cancel out after repeated application of this relation. The same is also true for the two remaining traces, which are given by

$$\mathrm{tr}(\mathbf{M\Sigma M'\Sigma}) = -\frac{8\sigma^4\rho^2}{F(1-\rho^2)(1+\rho^2)^3}[(1+\rho^2+\rho^4)DF - 1 + \rho^2 + \rho^6]$$

and

$$\mathrm{tr}(\mathbf{M'\Sigma M'\Sigma}) = \frac{8\sigma^4\rho^2}{F^2(1-\rho^2)^2(1+\rho^2)^4}[(1+2\rho^2+6\rho^4+2\rho^6+\rho^8)DF - 1 + \rho^2 - 4\rho^4 + 8\rho^6 + \rho^8 - \rho^{10}].$$

Then, for instance, the top left element of $\mathbf{J}(\sigma^2,\alpha) = \mathbb{E}[\mathrm{sc}_{\mathrm{C}}(\sigma^2,\alpha;\mathbf{y})\mathrm{sc}_{\mathrm{C}}(\sigma^2,\alpha;\mathbf{y})^T]$ can be calculated by

using (3.14), (3.17), (3.20) and Corollary 3.4 as follows:

$$\begin{aligned}\{\mathbf{J}(\sigma^2,\alpha)\}_{11} &= \mathbb{E}\left[\Big(-\frac{DF}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}\mathbf{y}\Big)^2\right]\\ &= \frac{(DF)^2}{4(\sigma^2)^2}-\frac{DF}{2(\sigma^2)^3}\mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma})+\frac{1}{4(\sigma^2)^4}\big(\mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma})^2+2\mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}\boldsymbol{\Sigma})\big)\\ &= \frac{1}{2\sigma^4(1+\rho^2)^2}[(1+4\rho^2+\rho^4)DF-2\rho^2+4\rho^4].\end{aligned}$$

From here onwards, to shorten the length of the expressions, we introduce the notation $(\mathbf{DF})^{(k)} \equiv ((DF)^k,(DF)^{k-1},...,(DF)^0)^T$. Then by following a similar procedure to the above to obtain the remaining elements, we can express $\mathbf{J}(\sigma^2,\alpha)$ as

$$\mathbf{J}(\sigma^2,\alpha)=\begin{bmatrix}\frac{1}{2\sigma^4(1+\rho^2)^2}\mathbf{j}_1^T(\mathbf{DF})^{(1)} & \frac{4\rho^2}{\sigma^2F(1-\rho^2)(1+\rho^2)^3}\mathbf{j}_2^T(\mathbf{DF})^{(1)}\\ \frac{4\rho^2}{\sigma^2F(1-\rho^2)(1+\rho^2)^3}\mathbf{j}_2^T(\mathbf{DF})^{(1)} & \frac{4\rho^2}{F^2(1-\rho^2)^2(1+\rho^2)^4}\mathbf{j}_3^T(\mathbf{DF})^{(1)}\end{bmatrix},$$

where

$$\mathbf{j}_1=\begin{bmatrix}1+4\rho^2+\rho^4\\ -2\rho^2+4\rho^4\end{bmatrix},\quad \mathbf{j}_2=\begin{bmatrix}1+\rho^2+\rho^4\\ -1+\rho^2+\rho^6\end{bmatrix},\quad \mathbf{j}_3=\begin{bmatrix}1+2\rho^2+6\rho^4+2\rho^6+\rho^8\\ -1+\rho^2-4\rho^4+8\rho^6+\rho^8-\rho^{10}\end{bmatrix}.$$

Next, we can express the inverse of $\mathbf{H}(\sigma^2,\alpha)$ from (3.18) as

$$\mathbf{H}(\sigma^2,\alpha)^{-1}=\frac{1}{\mathbf{r}^T(\mathbf{DF})^{(2)}}\begin{bmatrix}\frac{2\sigma^4}{1-\rho^2}\mathbf{h}^T(\mathbf{DF})^{(1)} & -2\sigma^2F(1+\rho^2)(DF-(1-\rho^2))\\ -2\sigma^2F(1+\rho^2)(DF-(1-\rho^2)) & \frac{F^2(1-\rho^2)(1+\rho^2)^2}{2\rho^2}DF\end{bmatrix},$$

where

$$\mathbf{h}=\begin{bmatrix}1+\rho^2+3\rho^4-\rho^6\\ -(1-\rho^2)(1+3\rho^4)\end{bmatrix},\quad \mathbf{r}=\begin{bmatrix}(1-\rho^2)^2\\ -1+8\rho^2-3\rho^4\\ -4\rho^2(1-\rho^2)\end{bmatrix}.$$

Finally, we obtain

$$\begin{aligned}\mathbf{G}(\sigma^2,\alpha)^{-1} &= \mathbf{H}(\sigma^2,\alpha)^{-1}\mathbf{J}(\sigma^2,\alpha)\mathbf{H}(\sigma^2,\alpha)^{-1} \\ &= \frac{1}{(\mathbf{r}^T(\mathbf{DF})^{(2)})^2}\begin{bmatrix} \frac{2\sigma^4}{1-\rho^2}\mathbf{g}_1^T(\mathbf{DF})^{(3)} & -2\sigma^2 F\mathbf{g}_2^T(\mathbf{DF})^{(3)} \\ -2\sigma^2 F\mathbf{g}_2^T(\mathbf{DF})^{(3)} & \frac{F^2(1-\rho^2)}{\rho^2}\mathbf{g}_3^T(\mathbf{DF})^{(3)} \end{bmatrix}, \end{aligned} \tag{3.21}$$

where

$$\mathbf{g}_1 = \begin{bmatrix} (1-\rho^2)^3(1+\rho^4) \\ 2(-1+8\rho^2-11\rho^4+15\rho^6-4\rho^8+\rho^{10}) \\ (1-\rho^2)(1-18\rho^2+26\rho^4-42\rho^6+\rho^8) \\ 2\rho^2(1-\rho^2)^2(3-5\rho^2+10\rho^4) \end{bmatrix}, \quad \mathbf{g}_2 = \begin{bmatrix} (1-\rho^2)^3 \\ -2+15\rho^2-17\rho^4+13\rho^6-\rho^8 \\ (1-\rho^2)(1-17\rho^2+13\rho^4-9\rho^6) \\ 2\rho^2(1-\rho^2)^2(3-2\rho^2) \end{bmatrix},$$

$$\mathbf{g}_3 = \begin{bmatrix} (1-\rho^2)^3 \\ -1+12\rho^2-16\rho^4+12\rho^6+\rho^8 \\ -2\rho^2(1-\rho^2)(3-8\rho^2+3\rho^4) \\ 4\rho^4(1-\rho^2)(-1+2\rho^2) \end{bmatrix}.$$

We can now compare (3.21) to the inverse Fisher information (3.5).

3.3.3 Asymptotics and Relative Efficiency

We first analyse the performance of the maximum composite conditional 2-nearest neighbours likelihood under the expanding domain framework. Note that each of the elements in (3.21) decreases at a rate of D^{-1}. Thus, we have the following approximation for $\mathbf{G}(\sigma^2,\alpha)^{-1}$:

$$\mathbf{G}(\sigma^2,\alpha)^{-1} \approx \begin{bmatrix} \frac{2\sigma^4(1+\rho^4)}{F(1-\rho^2)^2} & -\frac{2\sigma^2}{1-\rho^2} \\ -\frac{2\sigma^2}{1-\rho^2} & \frac{F}{\rho^2} \end{bmatrix}\frac{1}{D}.$$

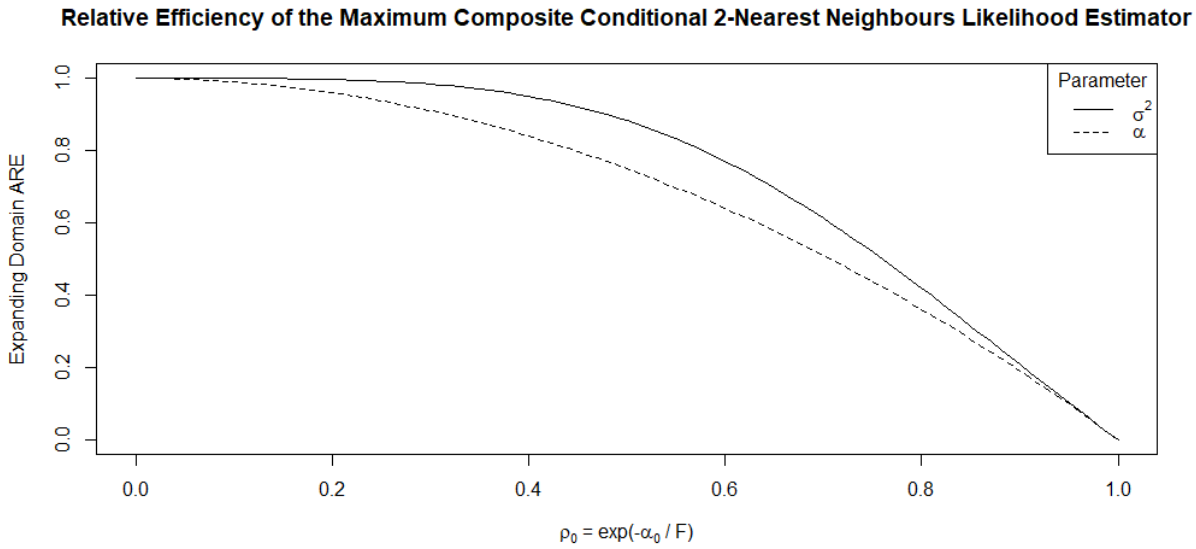

Figure 3.2: Expanding domain asymptotic relative efficiency of the maximum composite conditional 2-nearest neighbours likelihood estimator with respect to $\rho_0 = e^{-\frac{\alpha_0}{F}}$.

As a means of comparison, we can also obtain the corresponding large domain approximation for the inverse Fisher information matrix from (3.5):

$$\mathbf{I}(\sigma^2,\alpha)^{-1} \approx \begin{bmatrix} \frac{2\sigma^4(1+\rho^2)}{F(1-\rho^2)} & -2\sigma^2 \\ -2\sigma^2 & \frac{F(1-\rho^2)}{\rho^2} \end{bmatrix} \frac{1}{D}. \tag{3.22}$$

From this, the performance of the maximum composite likelihood estimator can be evaluated by calculating the expanding domain asymptotic relative efficiency (ARE) as follows:

$$\mathrm{ARE}(\hat{\sigma}^2_{\mathrm{CL}},\hat{\sigma}^2_{\mathrm{ML}}) = \lim_{D\to\infty} \frac{\mathrm{var}(\hat{\sigma}^2_{\mathrm{ML}})}{\mathrm{var}(\hat{\sigma}^2_{\mathrm{CL}})} = \lim_{D\to\infty} \frac{\{\mathbf{I}(\sigma_0^2,\alpha_0)^{-1}\}_{11}}{\{\mathbf{G}(\sigma_0^2,\alpha_0)^{-1}\}_{11}} = \frac{1-\rho_0^4}{1+\rho_0^4} = \frac{1-e^{-\frac{4\alpha_0}{F}}}{1+e^{-\frac{4\alpha_0}{F}}},$$

$$\mathrm{ARE}(\hat{\alpha}_{\mathrm{CL}},\hat{\alpha}_{\mathrm{ML}}) = \lim_{D\to\infty} \frac{\{\mathbf{I}(\sigma_0^2,\alpha_0)^{-1}\}_{22}}{\{\mathbf{G}(\sigma_0^2,\alpha_0)^{-1}\}_{22}} = 1-\rho_0^2 = 1-e^{-\frac{2\alpha_0}{F}}. \tag{3.23}$$

These two functions have been plotted against ρ_0 in Figure 3.2.

First, observe that the asymptotic relative efficiencies of both $\hat{\sigma}^2_{\mathrm{CL}}$ and $\hat{\alpha}_{\mathrm{CL}}$ are bounded between 0 and 1. This is to be expected as the maximum likelihood estimator asymptotically achieves the lowest possible

variance attainable by an estimator, as per Theorem 1.2.

However, of greater interest is the fact that both asymptotic relative efficiencies are decreasing functions of $\rho_0 \in (0,1)$; that is, the performance of the maximum composite likelihood estimators becomes worse as the strength of dependence between adjacent observations increases. Conversely, this means that the maximum composite conditional 2-nearest neighbours likelihood estimator becomes more favourable when the data are closer to being independent and identically distributed; that is, when the data are spaced far apart relative to the size of α. Another consequence of this result is that the maximum composite likelihood estimator performs very poorly under infill asymptotics where F tends to infinity. In fact, (3.21) actually increases to infinity under infill, so there is an inherent structural issue with this construction.

Overall, these results suggest that it may be better to construct a composite likelihood that approximates the full likelihood in structure such as the composite conditional K-sequential neighbours likelihood. This will allow its behaviour under the different asymptotic frameworks to be similar to the full likelihood.

3.4 Composite Marginal Blockwise Likelihood

Using the same approach as in Section 3.3 where we construct the composite likelihood and additively decompose the resulting composition matrix $\mathbf{M}$, the exact form of the sandwich covariance matrix for the composite marginal blockwise likelihood can also be derived. However, similar to the composite conditional K-sequential neighbours likelihood, there is some ambiguity in the construction of the likelihood. Hence, we will follow Caragea and Smith (2007) and consider a design where we partition the number line into B equal-sized blocks.

3.4.1 Construction of the Composite Likelihood

Suppose that the observations on the equally-spaced lattice in Figure 3.1 can be partitioned into B blocks containing (for convenience) the same number of observations W; that is, $N = DF = BW$ with $B, W \in \mathbb{Z}^+$. Now note that for $1 \leq b \leq B$, the joint distribution of $\mathbf{y}_b \equiv (y(s_{(b-1)W+1}), y(s_{(b-1)W+2}), ..., y(s_{bW}))^T$ is multivariate Gaussian with mean zero and covariance matrix $\mathbf{S}_0$, which is the same matrix as $\boldsymbol{\Sigma}$ from (3.2) except of size $W \times W$. Thus, the joint density can be written as

$$f(\mathbf{y}_b; \sigma^2, \alpha) = \frac{1}{(2\pi)^{\frac{W}{2}} \sigma^W (1-\rho^2)^{\frac{w-1}{2}}} \exp\Big(-\frac{1}{2\sigma^2}\mathbf{y}_b^T \mathbf{Q} \mathbf{y}_b\Big),$$

where $\rho = e^{-\frac{\alpha}{F}}$ as before in Section 3.3, and

$$\mathbf{Q} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & & & & & \\ -\rho & 1+\rho^2 & -\rho & & & & \\ & -\rho & 1+\rho^2 & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & 1+\rho^2 & -\rho & \\ & & & & -\rho & 1 & \end{bmatrix}.$$

Due to its tridiagonal structure, we may additively decompose $\mathbf{Q}$ as

$$\mathbf{Q} = \frac{1}{1-\rho^2}[\mathbf{I} + \rho^2 \mathbf{A} - \rho \mathbf{C}], \tag{3.24}$$

where the definitions of $\mathbf{A}$ and $\mathbf{C}$ are as per (3.12). The composite marginal blockwise log-likelihood is therefore

$$c\ell(\sigma^2, \alpha; \mathbf{y}) = \sum_{b=1}^{B} \log f(\mathbf{y}_b; \sigma^2, \alpha) = -\frac{DF}{2}\log\sigma^2 - \frac{DF-B}{2}\log\left(1-\rho^2\right) - \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}\mathbf{y} + \text{const.}, \tag{3.25}$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{Q} & & & \\ & \mathbf{Q} & & \\ & & \ddots & \\ & & & \mathbf{Q} \end{bmatrix}.$$

If $B = 1$, this is equivalent to the full log-likelihood in (3.1). Otherwise, the blockwise structure of $\mathbf{M}$ represents misspecified independence between blocks.

3.4.2 Derivation of the Sandwich Covariance Matrix

Using (3.25), the composite score function is given by

$$\text{sc}_{\text{C}}(\sigma^2, \alpha; \mathbf{y}) = \begin{bmatrix} -\frac{DF}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}\mathbf{y} \\ -\frac{(DF-B)\rho^2}{F(1-\rho^2)} - \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} \end{bmatrix}, \tag{3.26}$$

while the composite information function is given by

$$\text{info}_{\text{C}}(\sigma^2, \alpha; \mathbf{y}) = \begin{bmatrix} -\frac{DF}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\mathbf{y}^T\mathbf{M}\mathbf{y} & -\frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} \\ -\frac{1}{2(\sigma^2)^2}\mathbf{y}^T\mathbf{M}'\mathbf{y} & -\frac{2(DF-B)\rho^2}{F^2(1-\rho^2)^2} + \frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{M}''\mathbf{y} \end{bmatrix}, \tag{3.27}$$

where $\mathbf{M}'$ denotes differentiation of $\mathbf{M}$ with respect to α.

As in Section 3.2.2, we are interested in finding various traces of matrix products between $\mathbf{M}$ (and its derivatives) and $\mathbf{\Sigma}$ in order to derive the sandwich covariance matrix of the maximum composite marginal blockwise likelihood estimator. To aid in this procedure, we first additively decompose the relevant matrices. We let $\mathbf{M} = \mathbf{M}_{(1)} + \mathbf{M}_{(2)} + ... + \mathbf{M}_{(B)}$, where $\mathbf{M}_{(b)}$ is a $W \times W$ matrix containing only the b-th block of $\mathbf{M}$ (with all other elements set to zero). Also, we break down the structure of

$\mathbf{\Sigma} \in \mathbb{R}^{DF \times DF}$ into blocks of size $W \times W$ as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{S}_{(0)} & \mathbf{S}_{(1)} & \mathbf{S}_{(2)} & \dots & \mathbf{S}_{(B-2)} & \mathbf{S}_{(B-1)} \\ \mathbf{S}_{(-1)} & \mathbf{S}_{(0)} & \mathbf{S}_{(1)} & \dots & \mathbf{S}_{(B-3)} & \mathbf{S}_{(B-2)} \\ \mathbf{S}_{(-2)} & \mathbf{S}_{(-1)} & \mathbf{S}_{(0)} & \dots & \mathbf{S}_{(B-4)} & \mathbf{S}_{(B-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{S}_{(-(B-2))} & \mathbf{S}_{(-(B-3))} & \mathbf{S}_{(-(B-4))} & \dots & \mathbf{S}_{(0)} & \mathbf{S}_{(1)} \\ \mathbf{S}_{(-(B-1))} & \mathbf{S}_{(-(B-2))} & \mathbf{S}_{(-(B-3))} & \dots & \mathbf{S}_{(-1)} & \mathbf{S}_{(0)} \end{bmatrix},$$

where

$$\mathbf{S}_{(k)} = \sigma^2 \begin{bmatrix} \rho^{|Wk|} & \rho^{|Wk+1|} & \rho^{|Wk+2|} & \dots & \rho^{|W(k+1)-2|} & \rho^{|W(k+1)-1|} \\ \rho^{|Wk-1|} & \rho^{|Wk|} & \rho^{|Wk+1|} & \dots & \rho^{|W(k+1)-3|} & \rho^{|W(k+1)-2|} \\ \rho^{|Wk-2|} & \rho^{|Wk-1|} & \rho^{|Wk|} & \dots & \rho^{|W(k+1)-4|} & \rho^{|W(k+1)-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{|W(k-1)+2|} & \rho^{|W(k-1)+3|} & \rho^{|W(k-1)+4|} & \dots & \rho^{|Wk|} & \rho^{|Wk+1|} \\ \rho^{|W(k-1)+1|} & \rho^{|W(k-1)+2|} & \rho^{|W(k-1)+3|} & \dots & \rho^{|Wk-1|} & \rho^{|Wk|} \end{bmatrix}.$$

Then, for instance,

$$\operatorname{tr}(\mathbf{M}\mathbf{\Sigma}) = \sum_{b=1}^{B} \operatorname{tr}(\mathbf{M}_{(b)}\mathbf{\Sigma}) = \sum_{b=1}^{B} \operatorname{tr}(\mathbf{Q}\mathbf{S}_{(0)}) = B \operatorname{tr}(\mathbf{Q}\mathbf{S}_{(0)}).$$

The problem has now been reduced to a similar calculation as in (3.17). To accomplish this, we first make slight modifications to (3.16) to obtain

$$\operatorname{tr}(\mathbf{I}\mathbf{S}_{(0)}) = \sigma^2 W, \quad \operatorname{tr}(\mathbf{A}\mathbf{S}_{(0)}) = \sigma^2(W-2), \quad \operatorname{tr}(\mathbf{C}\mathbf{S}_{(0)}) = 2\sigma^2\rho(W-1).$$

Thus, by (3.24) and Lemma 3.1, we see that

$$\mathbb{E}[\mathbf{y}^T\mathbf{M}\mathbf{y}] = \operatorname{tr}(\mathbf{M}\mathbf{\Sigma}) = B \operatorname{tr}(\mathbf{Q}\mathbf{S}_{(0)})$$

$$= \frac{B}{1-\rho^2}\big(\mathrm{tr}(\mathbf{I}\boldsymbol{\Sigma}) + \rho^2\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) - \rho\,\mathrm{tr}(\mathbf{C}\boldsymbol{\Sigma})\big)$$
$$= \frac{\sigma^2 B}{1-\rho^2}\big(W + \rho^2(W-2) - 2\rho^2(W-1)\big)$$
$$= \sigma^2 DF. \tag{3.28}$$

Similarly, by noting from (3.24) that

$$\mathbf{Q}' = \frac{1}{F(1-\rho^2)^2}[-2\rho^2\mathbf{I} - 2\rho^2\mathbf{A} + (\rho^3+\rho)\mathbf{C}]$$

and

$$\mathbf{Q}'' = \frac{1}{F^2(1-\rho^2)^3}[4(\rho^4+\rho^2)\mathbf{I} + 4(\rho^4+\rho^2)\mathbf{A} - (\rho + 6\rho^3 + \rho^5)\mathbf{C}],$$

it can be shown that

$$\mathbb{E}[\mathbf{y}^T\mathbf{M}'\mathbf{y}] = -\sigma^2\frac{2\rho^2}{F(1-\rho^2)}(DF - B)$$

and

$$\mathbb{E}[\mathbf{y}^T\mathbf{M}''\mathbf{y}] = \sigma^2\frac{2\rho^2}{F^2(1-\rho^2)^2}(3+\rho^2)(DF-B).$$

Thus, by using (3.27), we obtain

$$\mathbf{H}(\sigma^2,\alpha) = \mathbb{E}[\mathrm{info}_{\mathrm{C}}(\sigma^2,\alpha;\mathbf{y})] = \begin{bmatrix} \frac{DF}{2(\sigma^2)^2} & \frac{\rho^2(DF-B)}{F\sigma^2(1-\rho^2)} \\ \frac{\rho^2(DF-B)}{F\sigma^2(1-\rho^2)} & \frac{\rho^2(1+\rho^2)(DF-B)}{F^2(1-\rho^2)^2} \end{bmatrix}. \tag{3.29}$$

Next, the task of deriving $\mathbf{J}(\sigma^2,\alpha) = \mathbb{E}[\mathrm{sc}_{\mathrm{C}}(\sigma^2,\alpha;\mathbf{y})\mathrm{sc}_{\mathrm{C}}(\sigma^2,\alpha;\mathbf{y})^T]$ requires finding the traces of the four-matrix products $\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}\boldsymbol{\Sigma}$, $\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'\boldsymbol{\Sigma}$ and $\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}'\boldsymbol{\Sigma}$. It is useful to observe that $\mathbf{S}_{(k)} = \rho^{W(k-1)}\mathbf{S}_{(1)}$ and $\mathbf{S}_{(-k)} = \rho^{W(k-1)}\mathbf{S}_{(-1)}$ for $k \geq 1$. Then, for instance,

$$\mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'\boldsymbol{\Sigma}) = \sum_{b=1}^{B}\sum_{c=1}^{B}\mathrm{tr}(\mathbf{M}_{(b)}\boldsymbol{\Sigma}\mathbf{M}'_{(c)}\boldsymbol{\Sigma}) = \sum_{b=1}^{B}\sum_{c=1}^{B}\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(c-b)}\mathbf{Q}'\mathbf{S}_{(b-c)})$$
$$= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \sum_{1\leq b<c\leq B}\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(c-b)}\mathbf{Q}'\mathbf{S}_{(b-c)}) + \sum_{1\leq c<b\leq B}\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(c-b)}\mathbf{Q}'\mathbf{S}_{(b-c)})$$

$$
\begin{aligned}
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \sum_{a=1}^{B-1}(B-a)\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(a)}\mathbf{Q}'\mathbf{S}_{(-a)}) + \sum_{a=1}^{B-1}(B-a)\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-a)}\mathbf{Q}'\mathbf{S}_{(a)}) \\
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)})\sum_{a=1}^{B-1}(B-a)\rho^{2W(a-1)} + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})\sum_{a=1}^{B-1}(B-a)\rho^{2W(a-1)} \\
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \big(\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})\big)\sum_{a=1}^{B-1}\sum_{k=1}^{B-a}(1)\rho^{2W(a-1)} \\
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \big(\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})\big)\sum_{k=1}^{B-1}\sum_{a=1}^{B-k}\rho^{2W(a-1)} \\
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \big(\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})\big)\sum_{k=1}^{B-1}\frac{1-\rho^{2W(B-k)}}{1-\rho^{2W}} \\
&= B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \frac{1}{1-\rho^{2W}}\Big(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\Big)\big(\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) + \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})\big). \qquad (3.30)
\end{aligned}
$$

Note that in the cases of $\mathrm{tr}(\mathbf{M}\Sigma\mathbf{M}\Sigma)$ and $\mathrm{tr}(\mathbf{M}\Sigma\mathbf{M}'\Sigma)$, we know from the cyclical invariance of the trace function that $\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}\mathbf{S}_{(-1)}) = \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}\mathbf{S}_{(1)})$ and $\mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) = \mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)})$. This allows us write

$$
\mathrm{tr}(\mathbf{M}\Sigma\mathbf{M}\Sigma) = B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}\mathbf{S}_{(0)}) + \frac{2}{1-\rho^{2W}}\Big(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\Big)\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}\mathbf{S}_{(-1)}) \qquad (3.31)
$$

and

$$
\mathrm{tr}(\mathbf{M}'\Sigma\mathbf{M}'\Sigma) = B\,\mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \frac{2}{1-\rho^{2W}}\Big(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\Big)\mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}).
$$

However, the same logic is not directly applicable to (3.30). In this case, we must first observe that $\mathbf{Q}$ and $\mathbf{Q}'$ are rotationally symmetric by 180 degrees; that is, for the antidiagonal identity matrix $\mathbf{R}$ (a square matrix with a diagonal of 1s from the top-right to the bottom-left), we have $\mathbf{Q} = \mathbf{R}\mathbf{Q}\mathbf{R}$ and $\mathbf{Q}' = \mathbf{R}\mathbf{Q}'\mathbf{R}$. Additionally, $\mathbf{S}_{(1)}$ and $\mathbf{S}_{(-1)}$ are 180-degree rotations of each other, such that $\mathbf{S}_{(1)} = \mathbf{R}\mathbf{S}_{(-1)}\mathbf{R}$. Thus,

$$
\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) = \mathrm{tr}((\mathbf{R}\mathbf{Q}\mathbf{R})\mathbf{S}_{(1)}(\mathbf{R}\mathbf{Q}'\mathbf{R})\mathbf{S}_{(-1)}) = \mathrm{tr}(\mathbf{Q}(\mathbf{R}\mathbf{S}_{(1)}\mathbf{R})\mathbf{Q}'(\mathbf{R}\mathbf{S}_{(-1)}\mathbf{R})) = \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(-1)}\mathbf{Q}'\mathbf{S}_{(1)}),
$$

and by applying this to (3.30), we obtain

$$
\mathrm{tr}(\mathbf{M}\Sigma\mathbf{M}'\Sigma) = B\,\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) + \frac{2}{1-\rho^{2W}}\Big(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\Big)\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}).
$$

The formulae for the traces of four-matrix products can then be found by first finding traces of the forms $\mathrm{tr}(\mathbf{U}\mathbf{S}_{(0)}\mathbf{V}\mathbf{S}_{(0)})$ and $\mathrm{tr}(\mathbf{U}\mathbf{S}_{(1)}\mathbf{V}\mathbf{S}_{(-1)})$, where $\mathbf{U}$ and $\mathbf{V}$ are various combinations of the three simple matrices $\mathbf{I}$, $\mathbf{A}$ and $\mathbf{C}$.

To begin, traces of the form $\mathrm{tr}(\mathbf{U}\mathbf{S}_{(0)}\mathbf{V}\mathbf{S}_{(0)})$ can be obtained from (3.19) with slight modifications as follows (and recalling that $u_n \equiv \sum_{k=1}^{n}(n-k)\rho^{2k}$):

$$\begin{aligned}
&\mathrm{tr}(\mathbf{I}\mathbf{S}_{(0)}\mathbf{I}\mathbf{S}_{(0)}) = \sigma^4[W+2u_W], && \mathrm{tr}(\mathbf{I}\mathbf{S}_{(0)}\mathbf{A}\mathbf{S}_{(0)}) = \sigma^4[W-2+2u_{W-1}],\\
&\mathrm{tr}(\mathbf{I}\mathbf{S}_{(0)}\mathbf{C}\mathbf{S}_{(0)}) = \sigma^4[4\rho(W-1)+4\rho u_{W-1}], && \mathrm{tr}(\mathbf{A}\mathbf{S}_{(0)}\mathbf{A}\mathbf{S}_{(0)}) = \sigma^4[W-2+2u_{W-2}],\\
&\mathrm{tr}(\mathbf{A}\mathbf{S}_{(0)}\mathbf{C}\mathbf{S}_{(0)}) = \sigma^4[4\rho(W-2)+4\rho u_{W-2}], && \mathrm{tr}(\mathbf{C}\mathbf{S}_{(0)}\mathbf{C}\mathbf{S}_{(0)}) = \sigma^4[2(1+\rho^2)(W-1)+8u_{W-1}].
\end{aligned} \tag{3.32}$$

Next, we shall define $v_n \equiv \sum_{k=1}^{2n-1}(n-|n-k|)\rho^{2k}$ and $q_n \equiv \sum_{k=1}^{n}\rho^{2k}$. Then, for instance,

$$\begin{aligned}
\mathrm{tr}(\mathbf{I}\mathbf{S}_{(1)}\mathbf{I}\mathbf{S}_{(-1)}) &= \mathrm{tr}(\mathbf{S}_{(1)}\mathbf{S}_{(-1)}) = \sum_{i=1}^{W}\sum_{j=1}^{W}\{\mathbf{S}_{(1)}\circ\mathbf{S}_{(-1)}^{T}\}_{ij}\\
&= \sum_{i=1}^{W}\sum_{j=1}^{W}\{\mathbf{S}_{(1)}\circ\mathbf{S}_{(1)}\}_{ij} = \sum_{i=1}^{W}\sum_{j=1}^{W}(\sigma^2\rho^{W+j-i})^2\\
&= \sigma^4\sum_{i=1}^{W}\sum_{k=1-i}^{W-i}\rho^{2(W+k)} = \sigma^4\sum_{k=1-W}^{W-1}(W-|k|)\rho^{2(W+k)}\\
&= \sigma^4 v_W.
\end{aligned}$$

In a similar manner, the remaining traces of the form $\mathrm{tr}(\mathbf{U}\mathbf{S}_{(0)}\mathbf{V}'\mathbf{S}_{(0)})$ can be shown to be

$$\begin{aligned}
&\mathrm{tr}(\mathbf{I}\mathbf{S}_{(1)}\mathbf{I}\mathbf{S}_{(-1)}) = \sigma^4 v_W, && \mathrm{tr}(\mathbf{I}\mathbf{S}_{(1)}\mathbf{A}\mathbf{S}_{(-1)}) = \sigma^4[\rho^2 v_{W-1} - \rho^{2W}],\\
&\mathrm{tr}(\mathbf{I}\mathbf{S}_{(1)}\mathbf{C}\mathbf{S}_{(-1)}) = \sigma^4[2\rho v_W - 2\rho^{2W-1}q_W], && \mathrm{tr}(\mathbf{A}\mathbf{S}_{(1)}\mathbf{A}\mathbf{S}_{(-1)}) = \sigma^4\rho^4 v_{W-2},\\
&\mathrm{tr}(\mathbf{A}\mathbf{S}_{(1)}\mathbf{C}\mathbf{S}_{(-1)}) = \sigma^4[2\rho^3 v_{W-1} - 2\rho^{2W-1}q_{W-1}], && \mathrm{tr}(\mathbf{C}\mathbf{S}_{(1)}\mathbf{C}\mathbf{S}_{(-1)}) = 4\sigma^4\rho^2 v_{W-1}.
\end{aligned} \tag{3.33}$$

Applying the results from (3.32) and (3.33), we can proceed with some lengthy algebra similar to (3.20) to obtain the following:

$$\begin{aligned}
&\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}\mathbf{S}_{(0)}) = \sigma^4 W, && \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) = -\tfrac{2\sigma^4\rho^2(W-1)}{F(1-\rho^2)},\\
&\mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(0)}\mathbf{Q}'\mathbf{S}_{(0)}) = \tfrac{2\sigma^4\rho^2(1+\rho^2)(W-1)}{F^2(1-\rho^2)^2}, && \mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}\mathbf{S}_{(-1)}) = \sigma^4\rho^2,\\
&\mathrm{tr}(\mathbf{Q}\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) = 0, && \mathrm{tr}(\mathbf{Q}'\mathbf{S}_{(1)}\mathbf{Q}'\mathbf{S}_{(-1)}) = 0.
\end{aligned} \tag{3.34}$$

For the traces involving $\mathbf{S}_{(0)}$, the algebra involves repeatedly applying $u_{n+1} = \rho^2(u_n + n)$ to cancel out the u_n terms, whereas for the traces involving $\mathbf{S}_{(1)}$ and $\mathbf{S}_{(-1)}$, we can use $v_{n+1} = v_n + 2\rho^{2n}q_n + \rho^{2(2n+1)}$ to cancel out the v_n terms.

We can then substitute the expressions from (3.34) into (3.31) and the corresponding versions for $\mathrm{tr}(\mathbf{M\Sigma M'\Sigma})$ and $\mathrm{tr}(\mathbf{M'\Sigma M'\Sigma})$ to find that

$$\mathrm{tr}(\mathbf{M\Sigma M\Sigma}) = \sigma^4\left[DF + \tfrac{2\rho^2}{1-\rho^{2W}}\left(B - \tfrac{1-\rho^{2DF}}{1-\rho^{2W}}\right)\right], \quad \mathrm{tr}(\mathbf{M\Sigma M'\Sigma}) = -\tfrac{2\sigma^4\rho^2(DF-B)}{F(1-\rho^2)},$$
$$\mathrm{tr}(\mathbf{M'\Sigma M'\Sigma}) = \tfrac{2\sigma^4\rho^2(1+\rho^2)(DF-B)}{F^2(1-\rho^2)^2}. \tag{3.35}$$

Then by using (3.26), (3.28), (3.35) and Corollary 3.4, we can calculate $\mathbf{J}(\sigma^2, \alpha)$ to be the following:

$$\mathbf{J}(\sigma^2,\alpha) = \mathbb{E}[\mathrm{sc}_\mathrm{C}(\sigma^2,\alpha;\mathbf{y})\mathrm{sc}_\mathrm{C}(\sigma^2,\alpha;\mathbf{y})^T] = \begin{bmatrix} \frac{1}{2\sigma^4}\left[DF + \frac{2\rho^2}{1-\rho^{2W}}\left(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\right)\right] & \frac{\rho^2}{\sigma^2F(1-\rho^2)}(DF-B) \\ \frac{\rho^2}{\sigma^2F(1-\rho^2)}(DF-B) & \frac{\rho^2(1+\rho^2)}{F^2(1-\rho^2)^2}(DF-B) \end{bmatrix}$$
$$= \mathbf{H}(\sigma^2,\alpha) + \begin{bmatrix} \frac{\rho^2}{\sigma^4(1-\rho^{2W})}\left(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\right) & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.36}$$

We observe from the above that $\mathbf{J}(\sigma^2, \alpha)$ differs from $\mathbf{H}(\sigma^2, \alpha)$ in (3.29) only in its top-left element. Additionally, this slight perturbation is zero if $B = 1$ (and $W = DF$), which reconciles with the fact that $\mathbf{J}(\sigma^2, \alpha) = \mathbf{H}(\sigma^2, \alpha)$ under the full likelihood.

Finally, by noting that the inverse of $\mathbf{H}(\sigma^2, \alpha)$ can be expressed as

$$\mathbf{H}(\sigma^2,\alpha)^{-1} = \frac{1}{(1-\rho^2)DF + 2\rho^2 B}\begin{bmatrix} 2(\sigma^2)^2(1+\rho^2) & -2\sigma^2F(1-\rho^2) \\ -2\sigma^2F(1-\rho^2) & F^2\frac{DF(1-\rho^2)^2}{(DF-B)\rho^2} \end{bmatrix},$$

we find that the sandwich variance $\mathbf{G}(\sigma^2, \alpha)^{-1}$ is given by

$$\mathbf{G}(\sigma^2,\alpha)^{-1} = \mathbf{H}(\sigma^2,\alpha)^{-1}\mathbf{J}(\sigma^2,\alpha)\mathbf{H}(\sigma^2,\alpha)^{-1}$$

$$= \mathbf{H}(\sigma^2, \alpha)^{-1} + \frac{4\rho^2}{(1-\rho^{2W})((1-\rho^2)DF + 2\rho^2 B)^2}\left(B - \frac{1-\rho^{2DF}}{1-\rho^{2W}}\right)\begin{bmatrix} (\sigma^2)^2(1+\rho^2)^2 & -\sigma^2 F(1-\rho^4) \\ -\sigma^2 F(1-\rho^4) & F^2(1-\rho^2)^2 \end{bmatrix}. \tag{3.37}$$

We can now draw comparisons between (3.37) and the inverse Fisher information in (3.5).

3.4.3 Asymptotics and Relative Efficiency

The composite marginal blockwise likelihood presents two main situations to consider as we increase our sample size to analyse asymptotic performance: keeping the number of blocks (B) fixed or keeping the block sizes (W) fixed.

In the case of a fixed B, we find under the expanding domain framework (where $D \to \infty$ and F is constant) that the second term in (3.37) decreases at the rate of $1/D^2$, and this is dominated by the $\mathbf{H}(\sigma^2, \alpha)^{-1}$ term that decreases at a rate of $1/D$. Hence, it is adequate to compare the asymptotic performance of the maximum composite blockwise likelihood estimator to the maximum likelihood estimator through $\mathbf{H}(\sigma^2, \alpha)^{-1}$ alone. Between the two estimation approaches, the only difference is in the value of B that is set; namely, the maximum likelihood estimator corresponds to the case where $B = 1$. However, we see that as $D \to \infty$, the small discrepancies caused by setting different values of B fall out quickly, so the asymptotic efficiencies of the maximum composite blockwise likelihood estimators for σ^2 and α relative to the corresponding maximum likelihood estimators are 1 for all choices of B. This is due to the individual blocks being able to grow in size to infinity, allowing each block to essentially mimic the structure of the full likelihood.

However, from a computational perspective, it is far more important to consider the case where the block sizes are fixed. If a closed-form expression for the determinant and inverse of the covariance matrix is unattainable, then they would need to be computed at a cost of $O((DF)^3)$. The composite blockwise likelihood reduces this cost to $O((DF)^3/B^2) = O(DFW^2)$ by requiring the determinant and inverse of

B matrices of size $W \times W$. Hence, the computational complexity of maximum composite blockwise likelihood estimation is better than that of maximum likelihood estimation only if B scales with the sample size. In fact, the best improvement in complexity arises if W is fixed, since estimation in this case has a linear computational cost.

When W is fixed and we let $B \to \infty$, the second term in $\mathbf{G}(\sigma^2, \alpha)^{-1}$ from (3.37) also decays at the rate of $1/D$ under expanding domain asymptotics, so it can no longer be ignored. Thus, we have the following approximation for $\mathbf{G}(\sigma^2, \alpha)^{-1}$ when D is large:

$$\mathbf{G}(\sigma^2,\alpha)^{-1} \approx \left(\frac{1}{F(1-\rho^2+2\rho^2/W)} \begin{bmatrix} 2(\sigma^2)^2(1+\rho^2) & -2\sigma^2 F(1-\rho^2) \\ -2\sigma^2 F(1-\rho^2) & F^2 \frac{W(1-\rho^2)^2}{(W-1)\rho^2} \end{bmatrix} + \frac{4\rho^2}{FW(1-\rho^{2W})(1-\rho^2+2\rho^2/W)^2} \begin{bmatrix} (\sigma^2)^2(1+\rho^2)^2 & -\sigma^2 F(1-\rho^4) \\ -\sigma^2 F(1-\rho^4) & F^2(1-\rho^2)^2 \end{bmatrix} \right) \frac{1}{D}.$$

We can then find the asymptotic efficiencies of the maximum composite blockwise likelihood estimators relative to the maximum likelihood estimators under the expanding domain framework by making use of the approximation for $\mathbf{I}(\sigma^2, \alpha)^{-1}$ from (3.22) to obtain

$$\mathrm{ARE}(\hat{\sigma}^2_{\mathrm{CL}}, \hat{\sigma}^2_{\mathrm{ML}}) = \lim_{D\to\infty} \frac{\{\mathbf{I}(\sigma_0^2, \alpha_0)^{-1}\}_{11}}{\{\mathbf{G}(\sigma_0^2, \alpha_0)^{-1}\}_{11}} = \frac{1-\rho_0^2+2\rho_0^2/W}{1-\rho_0^2} \Bigg/ \left(1 + \frac{2\rho_0^2(1+\rho_0^2)}{W(1-\rho_0^{2W})(1-\rho_0^2+2\rho_0^2/W)}\right),$$

$$\mathrm{ARE}(\hat{\alpha}_{\mathrm{CL}}, \hat{\alpha}_{\mathrm{ML}}) = \frac{1-\rho_0^2+2\rho_0^2/W}{1-\rho_0^2} \Bigg/ \left(\frac{W}{W-1} + \frac{4\rho_0^4}{W(1-\rho_0^{2W})(1-\rho_0^2+2\rho_0^2/W)}\right).$$

These two expressions have been plotted against W for different choices of the strength of dependence between adjacent observations $\rho_0 = \mathrm{e}^{-\frac{\alpha_0}{F}}$ in Figure 3.3.

First, we note that if $W = 1$, then maximum composite marginal blockwise likelihood estimation attains full asymptotic efficiency for σ^2 but zero efficiency for α. This is because $W = 1$ corresponds to a composite likelihood that simply takes the product of the univariate densities of each individual observation,

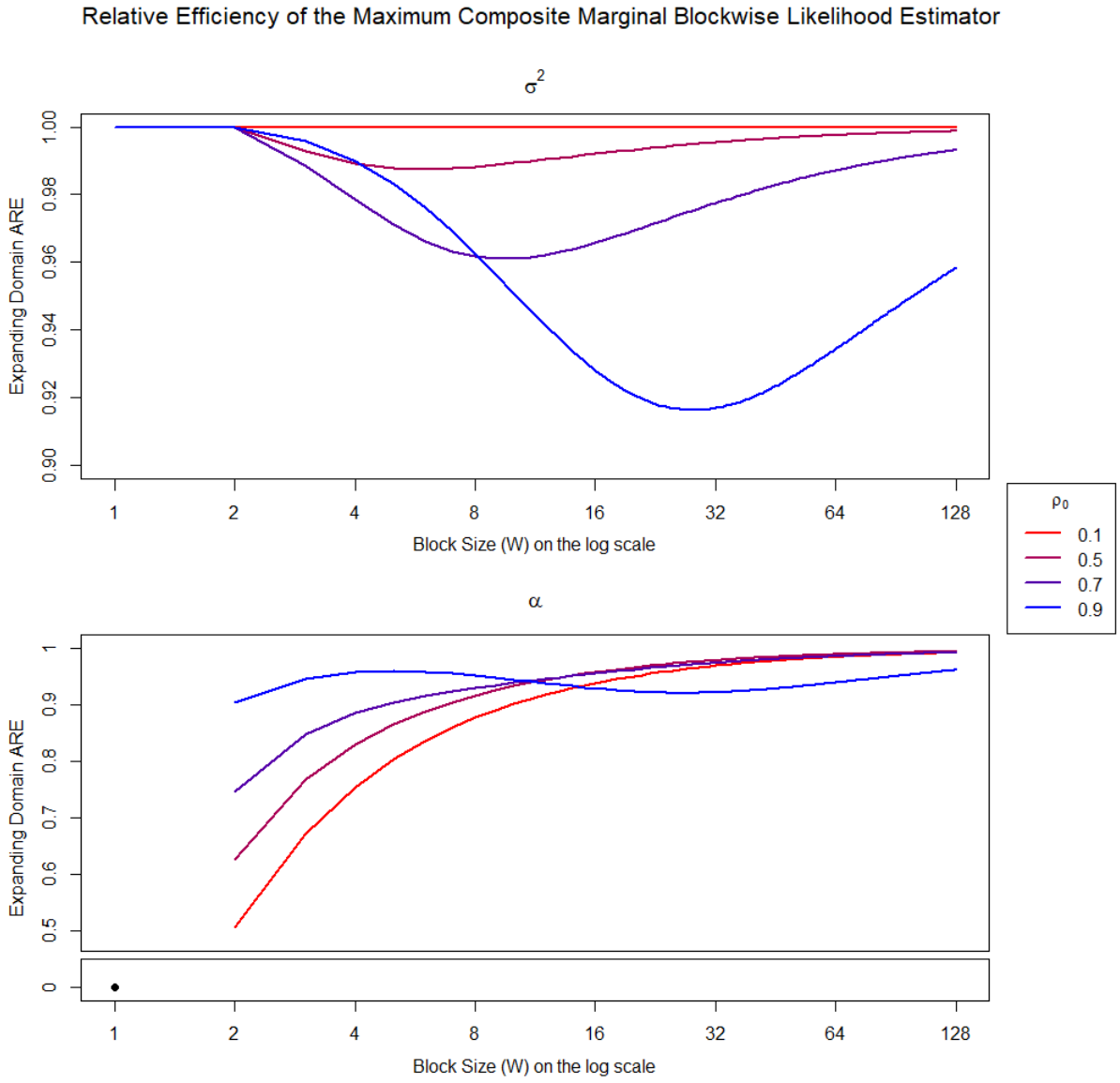

Figure 3.3: Expanding domain asymptotic relative efficiency of the maximum composite marginal blockwise likelihood estimator with respect to W for σ^2 (top) and α (bottom).

effectively treating the data as independent and identically distributed. As such, the composite likelihood does not contain α so it cannot be estimated, whereas all of the attention in estimation is focused on σ^2. In fact, the maximum composite likelihood estimator is simply $\hat{\sigma}^2_{\text{CL}} = \frac{\mathbf{y}^T\mathbf{y}}{DF}$.

Next, observe that the relative efficiency for σ^2 exhibits a peculiar decrease followed by an increase with respect to the block size. This is partially due to the fact that we have full asymptotic relative efficiency at both extremes: when $W = 1$ we have the i.i.d. composite likelihood, and when $W \to \infty$, each of the blocks are able to grow in size. A possible explanation for the loss of efficiency between these two

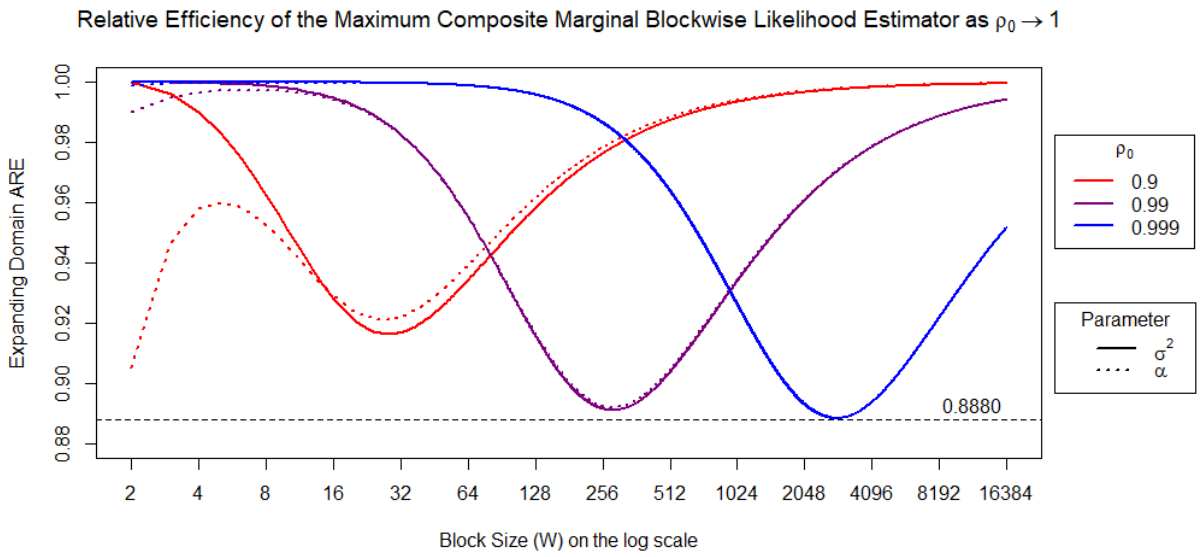

Figure 3.4: Expanding domain asymptotic relative efficiency of the maximum composite marginal blockwise likelihood estimator with respect to W for values of ρ_0 near 1.

extremes is that at the lower end, σ^2 is being treated as a "between blocks" variance parameter, whereas at the upper end, it is a 'within blocks" variance parameter. Hence, at the other values of W, σ^2 is a compromise between these two conflicting extremes.

When the strength of dependence is low (for example, $\rho_0 = 0.1$), we achieve close to full efficiency for σ^2 but mediocre efficiency for α for small values of W. This is attributable to the fact that more information about α can be acquired from observations that are weakly dependent than those that are strongly correlated. Hence, a greater amount of information is foregone by not accounting for the dependence between blocks when ρ_0 is low. In contrast, since the data are almost independent, little information is lost from using blocks when estimating the variance σ^2.

On the other hand, when the strength of dependence is high, the asymptotic relative efficiencies of both parameters exhibits peculiar behaviour. From Figure 3.4, observe for $W \geq 2$ that the relative efficiencies of the two parameters converge to the same value as $\rho_0 \to 1$. This can be verified algebraically as follows:

$$\lim_{\rho_0 \to 1^-} \frac{\mathrm{ARE}(\hat{\sigma}^2_{\mathrm{CL}}, \hat{\sigma}^2_{\mathrm{ML}})}{\mathrm{ARE}(\hat{\alpha}^2_{\mathrm{CL}}, \hat{\alpha}^2_{\mathrm{ML}})} = \lim_{\rho_0 \to 1^-} \frac{W^2(1-\rho_0^{2W})(1-\rho_0^2+2\rho_0^2/W)/(W-1)+4\rho_0^4}{W(1-\rho_0^{2W})(1-\rho_0^2+2\rho_0^2/W)+2\rho_0^2(1+\rho_0^2)}$$

$$= \lim_{\rho_0 \to 1^-} \frac{4\rho_0^4}{2\rho_0^2(1+\rho_0^2)} = 1.$$

Furthermore, the minimum of the curve is attained at a larger block size and lower relative efficiency; though it can be shown numerically that the lowest relative efficiency for any W and $\rho_0 \in [0,1)$ is 0.8880 (to four decimal places). Due to this right-shifting behaviour, under hybrid asymptotics where $\rho_0 \to 1$, we achieve full efficiency for both σ^2 and α for $W \geq 2$.

Given the various non-trivial relationships between the block size, the strength of dependence and the relative efficiencies of the estimators for both σ^2 and α, there is no straightforward optimal choice of block size. However, we would want to only consider small block sizes in order to benefit the most from a computational standpoint. Since the efficiency for σ^2 can fall no lower than 0.8880, which is still quite high, it would be reasonable to make a choice for W based solely on the efficiency for α in this situation. An approach to do this would be to specify a desired level of relative efficiency for α and solve for W, which would be found numerically. However, if we have reason to believe that the data are quite weakly dependent ($\rho_0 \approx 0$), then we note that $\text{ARE}(\hat{\alpha}_{\text{CL}}, \hat{\alpha}_{\text{ML}}) \approx (W-1)/W$, which serves as a worst-case scenario. Our choice of W based on a desired level of relative efficiency q could then be $W = \lceil 1/(1-q) \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

Chapter 4

Data and Simulation: Gaussian Exponential Covariance Model in Two Dimensions

The theoretical derivations in Chapter 3 are obtained in a simple one-dimensional framework where observations are equally-spaced and have no measurement error. However, in most applications with geostatistical data, this is unrealistic. In this chapter, we will explore the statistical and computational efficiency of maximum composite likelihood estimation in a more practical two-dimensional setting.

We will first analyse data for the average maximum temperature of the United States during January 2000 in Section 4.1, allowing us to identify an appropriate Gaussian spatial regression model. An iterative scheme will then be highlighted in Section 4.2 to estimate the model parameters using maximum likelihood and maximum composite likelihood estimation. Due to the irregular spacing of the locations at which temperatures are recorded, we will address the computational implementation of various choices of composite likelihood. In order to quantify the uncertainty around the estimated parameters, we will derive an expression for the sandwich covariance matrix in Section 4.3 that can be implemented in a computationally straightforward manner.

In Section 4.4, we will then use the results from Sections 4.2 and 4.3 to compare various choices of composite likelihood from a statistical and computational standpoint when applied to the selected spatial regression model for our data. Finally, in Section 4.5, we will use this model to design a simulation study for investigating the statistical performance of maximum composite likelihood estimation relative to maximum likelihood estimation.

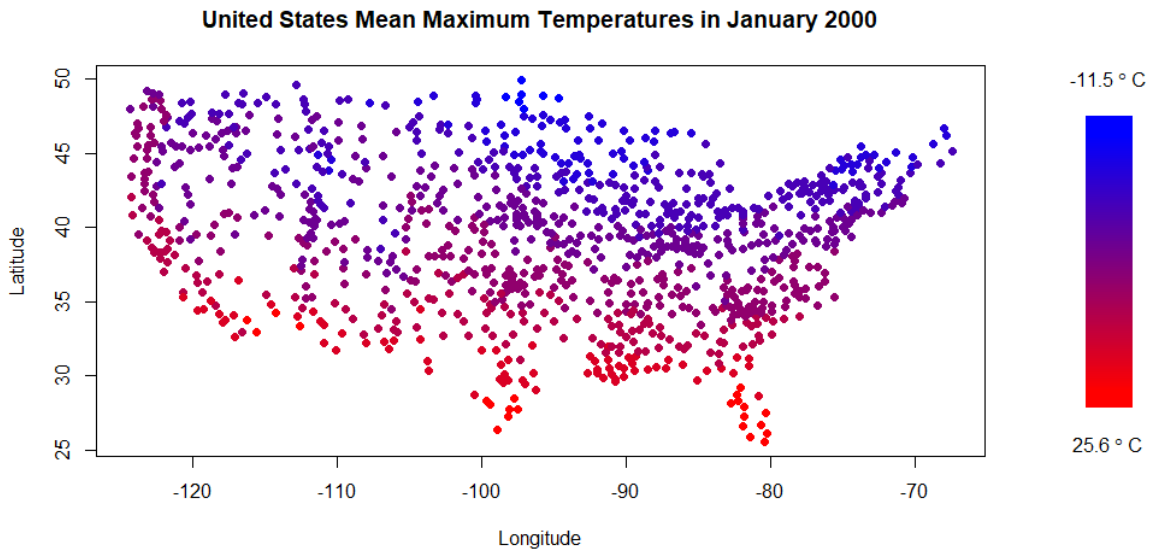

Figure 4.1: United States mean maximum temperatures in January 2000 at $N = 1052$ locations.

4.1 Analysis of Maximum Temperature Dataset

In order to draw similarities with our work in the one-dimensional setting, we have chosen to investigate temperature data as it is a continuous response variable that can be modelled generally well using a Gaussian distribution. This dataset was obtained from the National Oceanic and Atmospheric Administration, and is available at https://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.GHCN/.v2/.adjusted/.Max/.temp/ (Peterson and Vose, 1997). The full dataset includes mean monthly maximum temperatures since 1851, which have been adjusted to account for changes in data collection apparatus over time. However, for illustrative purposes, we have chosen only to model data for the month of January 2000 in order to keep the focus on spatial variability. The data have been collected at many stations globally, but since there are relatively few observations outside of the United States, we have only considered the $N = 1052$ observations that lie in the United States. A graphical representation of the locations of these stations, as well as temperatures recorded, is presented in Figure 4.1.

It is clear from this plot that mean maximum temperatures are generally higher for locations with a lower `latitude`, which is due to their close proximity to the equator. Hence, we have fitted a spatial

linear regression model to account for this non-constant mean level. Based on the variables `longitude` and `latitude` that are in the dataset, we have chosen to include the covariates `longitude`, `latitude` and $\texttt{latitude}^2$; all of which were found to be statistically significant (p-values $\ll 0.05$) in a multiple linear regression model. While other higher order terms and interactions between `longitude` and `latitude` were also significant in this particular month of data, we have chosen only to include these three covariates as they were found to be consistently significant even in other months that we tested.

Based on the linear model fit, we then used the mean-normalised data $\{u_1, ..., u_N\}$, as obtained from the raw residuals, to test for a spatial dependence structure. A common initial test for this uses Moran's I (see Cressie and Wikle (2011, p. 167-168)), which is given by $I = \frac{N}{\sum_{i=1}^N \sum_{j=1}^N w_{ij}} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} u_i u_j}{\sum_{i=1}^N u_i^2}$ where the weights w_{ij} are appropriately chosen to describe the supposed strength of spatial dependence between $\mathbf{s}_i$ and $\mathbf{s}_j$ for $i \neq j$ (and are zero for $i = j$). Using this, we can conduct a hypothesis test where a value of I far different from $\mathbb{E}_{H_0}[I] = -\frac{1}{N-1}$ equates to rejecting the null hypothesis of spatial independence. The standard normal pivot is usually used for this purpose; see Moran (1950) for the expression of $\text{var}_{H_0}[I]$.

For our dataset, we have calculated Moran's I using weights corresponding to the inverse of the pseudo-distance between two observations; which is calculated as $\sqrt{\texttt{longitude}^2 + \texttt{latitude}^2}$. From this, we obtained a p-value of effectively zero, strongly suggesting that there was still spatial dependence in our mean-normalised data.

In order to determine an appropriate stationary covariance structure for the data, we used variograms, as described in Section 1.2. First, directional semivariograms were obtained to identify whether there were any major differences in the decay of dependence in different directions. From the middle plot in Figure 4.2, we see that the directional semivariograms roughly overlap, so it is reasonable to assume an isotropic covariance structure. In this situation, it is preferable to use Haversine (great-circle) distances as they are a more accurate measure of distance between two observations on the surface of the Earth (Sinnott, 1984). Based on the right plot in Figure 4.2, we see that the omnidirectional semivariogram initially begins at a non-zero level, which suggests the presence of a nugget effect (measurement error).

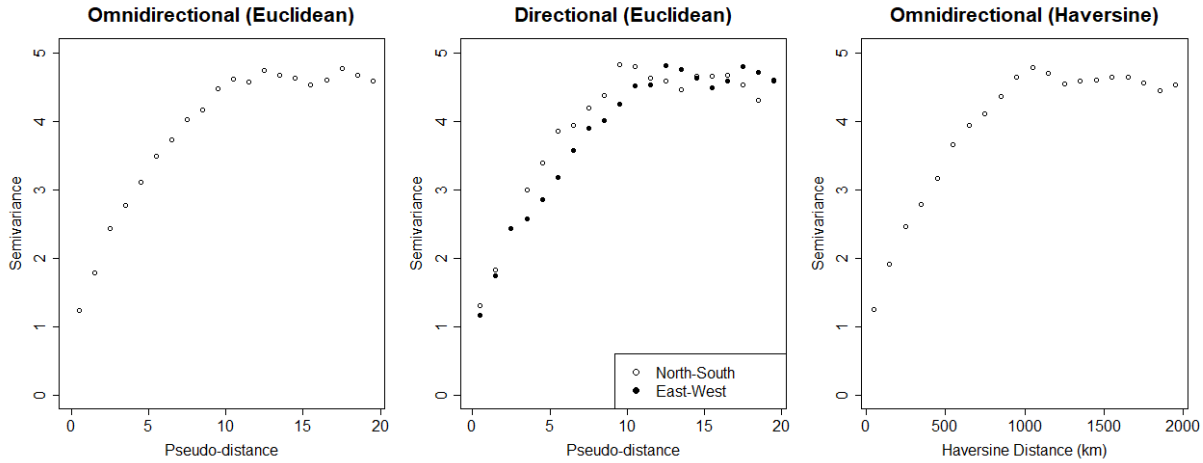

Figure 4.2: Empirical semivariograms of the mean-normalised maximum temperature data. All three semivariograms exhibit a similar non-zero intercept and plateau.

As the separation distance increases, the empirical semivariance increases before reaching a plateau at a Haversine distance of approximately 1000 kilometres. These features can also be observed in the semivariograms based on Euclidean distance. Although there are many covariance structures that would have these semivariogram characteristics, we ended up choosing an isotropic exponential covariance structure to draw comparisons with Chapter 3.

Overall, we have chosen to fit the spatial regression model $\mathbf{z} \sim N(\mathbf{X}\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ with the regression component incorporating an intercept, `longitude`, `latitude` and `latitude`^{2}, and covariance structure $\boldsymbol{\Sigma}_{0,ij} = \tau_0^2 I(i=j) + \sigma_0^2 \exp(-\alpha_0 \|\mathbf{s}_i - \mathbf{s}_j\|)$, where $\|\cdot\|$ corresponds to the Haversine distance (in thousands of kilometres). Thus, our model has the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T)^T$, where $\boldsymbol{\phi} = (\sigma^2, \alpha, \tau^2)^T$.

4.2 Maximum Composite Likelihood Estimation

For our analysis, we have considered the full likelihood and the three types of composite likelihood as specified in Definitions 2.1, 2.2 and 2.3. The implementation of the composite conditional K-nearest neighbours likelihood is straightforward as we simply need to identify the K closest observations to each individual observation. On the other hand, the composite conditional K-sequential neighbours likelihood is dependent on the order of observations. For simplicity, we chose a starting point near the

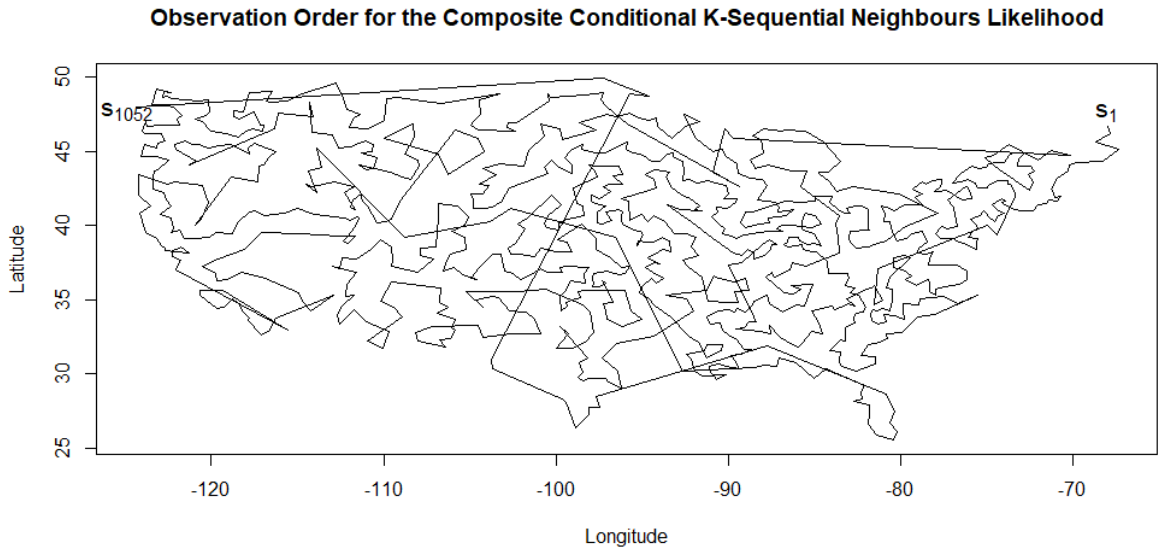

Figure 4.3: An example setup of the order in which observations in the composite conditional K-sequential neighbours likelihood are to be included. The starting point was chosen to the location at the top right of the United States as it is rather isolated.

North-Eastern border of the United States and successively included the location of the closest station that had yet to be selected. As illustrated in Figure 4.3, due to the irregular spacing of the stations, this often leads to sections where large jumps occur when the sequence of locations ends up at a "dead end", but such a setup is sufficient for the purposes of investigating the effect of increasing K.

There is also some ambiguity in the selection of blocks in the composite marginal blockwise likelihood. In line with our choice of blocks in Section 3.4, we have constructed blocks that involve observations close to each other. A fast approach for implementing this is to use hierarchical clustering to group observations based on Haversine distances. Although block sizes are unlikely to be equal, this comes with the advantage of having a nested block scheme when comparing the composite likelihood for different numbers of blocks. We also found that agglomerative clustering with complete linkage produced block that were closer in size than other linkage methods.

In order to account for the structural differences of the various composite likelihood functions, we propose a general method that allows for the estimation of parameters in the Gaussian spatial regression model. This is a straightforward extension of the iterative estimation algorithm often used in spa-

tial regression (Mardia and Marshall, 1984). Recall from Theorem 1.4 that both composite marginal and composite conditional likelihoods can be broken down into some number of marginal densities B, each involving a block of observations $\mathbf{z}_b$. Hence, we can express any composite log-likelihood as $c\ell(\boldsymbol{\theta};\mathbf{z}) = \sum_{b=1}^{B} m_b \log f(\mathbf{z}_b;\boldsymbol{\theta})$ with $m_b \in \{-1,1\}$, where $m_b = -1$ if the b-th block of observations is conditioned on in a composite conditional likelihood. Now since we have $\mathbf{z}_b \sim N(\mathbf{X}_b\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{0,b})$ for a Gaussian spatial regression model, the composite log-likelihood is as follows:

$$c\ell(\boldsymbol{\theta};\mathbf{z}) = \sum_{b=1}^{B} m_b \left[-\frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \frac{1}{2}(\mathbf{z}_b - \mathbf{X}_b\boldsymbol{\beta})^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{z}_b - \mathbf{X}_b\boldsymbol{\beta}) \right]. \tag{4.1}$$

If we differentiate (4.1) with respect to $\boldsymbol{\beta}$ and set it equal to zero, we find that

$$\begin{aligned} 0 &\overset{\text{set}}{=} \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\theta};\mathbf{z}) = -\sum_{b=1}^{B} m_b \mathbf{X}_b^T \boldsymbol{\Sigma}_b^{-1}(\mathbf{z}_b - \mathbf{X}_b\boldsymbol{\beta}) \\ &\implies \boldsymbol{\beta} = \left\{ \sum_{b=1}^{B} m_b \mathbf{X}_b^T \boldsymbol{\Sigma}_b^{-1} \mathbf{X}_b \right\}^{-1} \sum_{b=1}^{B} m_b \mathbf{X}_b^T \boldsymbol{\Sigma}_b^{-1} \mathbf{z}_b. \end{aligned}$$

We can therefore iterate between updating $\boldsymbol{\beta}$ given the other parameters $\boldsymbol{\phi} = (\sigma^2, \alpha, \tau^2)^T$ in a generalised least squares step, and then updating each $\boldsymbol{\Sigma}_b$ (which are functions of $\boldsymbol{\phi}$) given $\boldsymbol{\beta}$. For the full likelihood, Mardia and Marshall (1984) used Fisher scoring to update $\boldsymbol{\phi}$, and so analogous to the Fisher information matrix $I_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathbb{E}[-\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}\ell(\theta;\mathbf{y})]$, we can follow the suggestion of Huang and Ferrari (2017) and use $H_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathbb{E}[-\frac{\partial^2}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}c\ell(\theta;\mathbf{y})]$ for updating the spatial parameters in all of the composite likelihoods. To summarise, for an appropriate starting value $\boldsymbol{\phi}^{(0)}$, we can iterate between the following until convergence:

$$\begin{cases} \hat{\boldsymbol{\beta}}^{(k+1)} = \left\{ \sum_{b=1}^{B} m_b \mathbf{X}_b^T \{\hat{\boldsymbol{\Sigma}}_b^{-1}\}^{(k)} \mathbf{X}_b \right\}^{-1} \sum_{b=1}^{B} m_b \mathbf{X}_b^T \{\hat{\boldsymbol{\Sigma}}_b^{-1}\}^{(k)} \mathbf{z}_b, \\ \hat{\boldsymbol{\phi}}^{(k+1)} = \hat{\boldsymbol{\phi}}^{(k)} + \mathbf{H}_{\boldsymbol{\phi}}(\boldsymbol{\theta})^{-1} \frac{\partial}{\partial\boldsymbol{\phi}} c\ell(\boldsymbol{\theta};\mathbf{z})\Big|_{\boldsymbol{\theta}=(\{\hat{\boldsymbol{\beta}}^{(k+1)}\}^T, \{\hat{\boldsymbol{\phi}}^{(k)}\}^T)^T}. \end{cases} \tag{4.2}$$

Since the algorithm requires $H_{\boldsymbol{\phi}}(\boldsymbol{\theta})$, we will need to derive an expression for this. This will be shown in the next section as it is a component of the sandwich information matrix $G_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ which will be used for variance estimation.

4.3 Variance Estimation of Maximum Composite Likelihood Estimates

Under the assumption that we have approximate normality of the maximum likelihood or maximum composite likelihood estimator, a common method of producing confidence intervals or conducting hypothesis tests on parameters is to use the Wald statistic. This is where we make use of the approximate standard normal pivot $\frac{\hat{\theta}_j - \theta_{0,j}}{\sqrt{\text{var}[\hat{\theta}_j]}} \dot{\sim} N(0,1)$ for a given parameter θ_j. The challenge is then to estimate var$[\hat{\theta}]$, which is taken to be the j-th diagonal element of an estimate of $\mathbf{I}(\boldsymbol{\theta}_0)^{-1}$ or $\mathbf{G}(\boldsymbol{\theta}_0)^{-1} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$. This is typically done either by obtaining an exact expression for the sandwich covariance matrix and then plugging in the maximum composite likelihood estimator so that $\hat{\text{var}}[\hat{\theta}] = \mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{CL}})^{-1}$, or using a sample-based estimate of $\mathbf{G}(\boldsymbol{\theta}_0)$, which is denoted as $\hat{\mathbf{G}}(\hat{\boldsymbol{\theta}}_{\text{CL}})$.

Although more computationally efficient than computing $\mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{CL}})$, sample-based estimation introduces a further level of uncertainty in variance estimation. For both methods, the common issue is in estimating $\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}[\text{sc}_{\text{C}}(\boldsymbol{\theta};\mathbf{z})\text{sc}_{\text{C}}(\boldsymbol{\theta};\mathbf{z})^T]$, which involves fourth-order moments. In a geostatistical setting, Heagerty and Lele (1998) proposed the use of a window subsampling method, where $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_{\text{CL}}) = \frac{1}{S}\sum_{s=1}^{S} \text{dim}(\mathbf{z}_s)\text{sc}_{\text{C}}(\hat{\boldsymbol{\theta}}_{\text{CL}};\mathbf{z}_s)\{\text{sc}_{\text{C}}(\hat{\boldsymbol{\theta}}_{\text{CL}};\mathbf{z}_s)\}^T$ for S possibly overlapping subsets of $\mathbf{z}$. This is a consistent estimator of the true asymptotic variance, but its precision will vary depending on the selection of windows. Thus, given that our sample size is manageable enough to compute $\mathbf{J}(\hat{\boldsymbol{\theta}}_{\text{CL}})$ and this is the preferred means of estimating variance where possible (Stein et al., 2004), we will proceed by deriving a general expression for the sandwich covariance matrix.

In the context of Gaussian spatial regression, an exact expression for $\mathbf{I}(\boldsymbol{\theta})$ is available due to Mardia and Marshall (1984). We shall apply some of their intermediate calculations in order to derive expressions for $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ that work for any choice of composite likelihood. First, in the full likelihood case, the log-likelihood is as follows:

$$\ell(\theta;\mathbf{z}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}).$$

The first-order partial derivatives of the log-likelihood are given by

$$\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\theta;\mathbf{z}) = -\mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}),$$

and for each of the p parameters in $\boldsymbol{\phi}$,

$$\frac{\partial}{\partial \phi_j}\ell(\theta;\mathbf{z}) = -\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j}\Big) - \frac{1}{2}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^T\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \phi_j}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}).$$

Note that we can alternatively write $\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \phi_j} = -\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j}\boldsymbol{\Sigma}^{-1}$, which allows us to avoid having to explicitly obtain partial derivatives of the elements of $\boldsymbol{\Sigma}^{-1}$. Mardia and Marshall (1984) showed that the Fisher information matrix is block-diagonal and can be expressed as $\mathbf{I}(\boldsymbol{\theta}) = \mathrm{diag}(\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \mathbf{I}_{\boldsymbol{\phi}}(\boldsymbol{\theta}))$, where $\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbb{E}[-\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\ell(\theta;\mathbf{z})] = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\{\mathbf{I}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\}_{ij} = \mathbb{E}[-\frac{\partial^2}{\partial \phi_i \partial \phi_j}\ell(\theta;\mathbf{z})] = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j})$.

Now for a composite likelihood, by using (4.1), we find that $\mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}[-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}c\ell(\theta;\mathbf{z})]$ also has a block-diagonal structure, where

$$\mathbf{H}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \sum_{b=1}^{B} m_b\mathbf{X}_b^T\boldsymbol{\Sigma}_b^{-1}\mathbf{X}_b, \tag{4.3}$$

and

$$\{\mathbf{H}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\}_{ij} = \frac{1}{2}\sum_{b=1}^{B} m_b\mathrm{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_j}\Big). \tag{4.4}$$

Next, to derive $\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}[\mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z})\mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{z})^T]$, note that

$$\frac{\partial}{\partial \boldsymbol{\beta}}c\ell(\theta;\mathbf{z}) = -\sum_{b=1}^{B} m_b\mathbf{X}_b^T\boldsymbol{\Sigma}_b^{-1}(\mathbf{z}_b-\mathbf{X}_b\boldsymbol{\beta})$$

as before, and

$$\frac{\partial}{\partial \phi_j}\ell(\theta;\mathbf{z}) = \sum_{b=1}^{B} m_b\left[-\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_j}\Big) - \frac{1}{2}(\mathbf{z}_b-\mathbf{X}_b\boldsymbol{\beta})^T\frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_j}(\mathbf{z}_b-\mathbf{X}_b\boldsymbol{\beta})\right].$$

We will break down $\mathbf{J}(\boldsymbol{\theta})$ as follows and consider each component separately:

$$\mathbf{J}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) & \mathbf{J}_{\boldsymbol{\beta}:\boldsymbol{\phi}}(\boldsymbol{\theta}) \\ \mathbf{J}_{\boldsymbol{\beta}:\boldsymbol{\phi}}(\boldsymbol{\theta})^T & \mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \end{bmatrix}.$$

To begin, we note that $\mathbf{J}_{\boldsymbol{\beta}:\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathbf{0}$ since the expectation of the product of any odd number of zero-mean Gaussian random variables is zero (Isserlis, 1918). Hence, $\mathbf{J}(\boldsymbol{\theta})$ is also block-diagonal and we can write $\mathbf{J}(\boldsymbol{\theta}) = \text{diag}(\mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta}))$. Next, we have that

$$\begin{aligned} \mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= \sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \mathbf{X}_b^T \boldsymbol{\Sigma}_b^{-1} \mathbb{E}[(\mathbf{z}_b - \mathbf{X}_b\boldsymbol{\beta})(\mathbf{z}_c - \mathbf{X}_c\boldsymbol{\beta})^T](\boldsymbol{\Sigma}_c^{-1})^T \mathbf{X}_c \\ &\equiv \sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \mathbf{X}_b^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{\Sigma}_{b:c} \boldsymbol{\Sigma}_c^{-1} \mathbf{X}_c. \end{aligned} \tag{4.5}$$

Finally, recall Lemma 3.1 and Corollary 3.4, which are formulae for the expectation of a quadratic and quartic form, respectively. Also note from before that $\frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_i} = -\boldsymbol{\Sigma}_b^{-1} \frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i} \boldsymbol{\Sigma}^{-1}$. To simplify the notation, let $\mathbf{u}_b = \mathbf{z}_b - \mathbf{X}_b\boldsymbol{\beta}$. Thus, we obtain

$$\begin{aligned} \{\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{\theta})\}_{ij} &= \frac{1}{4}\sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \Bigg(\text{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\Big)\text{tr}\Big(\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \phi_j}\Big) + \text{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\Big)\mathbb{E}\Big[\mathbf{u}_c^T \frac{\partial \boldsymbol{\Sigma}_c^{-1}}{\partial \phi_j}\mathbf{u}_c\Big] \\ &\qquad + \text{tr}\Big(\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \phi_j}\Big)\mathbb{E}\Big[\mathbf{u}_b^T \frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_i}\mathbf{u}_b\Big] + \mathbb{E}\Big[\mathbf{u}_b^T \frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_i}\mathbf{u}_b\mathbf{u}_c^T \frac{\partial \boldsymbol{\Sigma}_c^{-1}}{\partial \phi_j}\mathbf{u}_c\Big] \Bigg) \\ &= \frac{1}{4}\sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \Bigg(-\text{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\Big)\text{tr}\Big(\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \phi_j}\Big) \\ &\qquad + \mathbb{E}\left[\begin{bmatrix}\mathbf{u}_b^T & \mathbf{u}_c^T\end{bmatrix}\begin{bmatrix}\frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{u}_b \\ \mathbf{u}_c\end{bmatrix}\begin{bmatrix}\mathbf{u}_b^T & \mathbf{u}_c^T\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\Sigma}_c^{-1}}{\partial \phi_j}\end{bmatrix}\begin{bmatrix}\mathbf{u}_b \\ \mathbf{u}_c\end{bmatrix}\right] \Bigg) \\ &= \frac{1}{2}\sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \,\text{tr}\left(\begin{bmatrix}\frac{\partial \boldsymbol{\Sigma}_b^{-1}}{\partial \phi_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_b & \boldsymbol{\Sigma}_{b:c} \\ (\boldsymbol{\Sigma}_{b:c})^T & \boldsymbol{\Sigma}_c\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\Sigma}_c^{-1}}{\partial \phi_j}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_b & \boldsymbol{\Sigma}_{b:c} \\ (\boldsymbol{\Sigma}_{b:c})^T & \boldsymbol{\Sigma}_c\end{bmatrix}\right) \\ &= \frac{1}{2}\sum_{b=1}^{B}\sum_{c=1}^{B} m_b m_c \,\text{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_{b:c}\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \phi_j}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\Sigma}_{b:c})^T\Big). \end{aligned} \tag{4.6}$$

From a computational standpoint, (4.6) is expressed in a way that only requires the manual implementation of the covariance matrix and its first-order partial derivatives. In most practical situations, $\boldsymbol{\Sigma}_b^{-1}$

does not have a simple form, and so it will be left to computation. Fortunately, if the block sizes of the terms in the composite likelihood are small, this will not be a costly computation.

An alternative way to express (4.6) is as follows:

$$\{\mathbf{J}_{\phi}(\boldsymbol{\theta})\}_{ij} = \frac{1}{2}\sum_{b=1}^{B} \mathrm{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_j}\Big) + \frac{1}{2}\sum_{b\neq c}^{B} m_b m_c \,\mathrm{tr}\Big(\boldsymbol{\Sigma}_b^{-1}\frac{\partial \boldsymbol{\Sigma}_b}{\partial \phi_i}\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_{b:c}\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \phi_j}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\Sigma}_{b:c})^T\Big).$$

In this form, we see that the first term is equivalent to $\{\mathbf{H}_{\phi}(\boldsymbol{\theta})\}_{ij}$ from (4.4) if $m_b = 1$ for all $b = 1,2,...,B$; that is, if we have a composite marginal likelihood. This is consistent with the form of $\mathbf{J}(\boldsymbol{\theta})$ from (3.36) that was derived in Chapter 3 for the composite marginal blockwise likelihood.

4.4 Application to Maximum Temperature Dataset

To begin our analysis of the spatial regression model for the maximum temperature data, we have looked at residual diagnostics. Due to the spatial covariance structure, we have standardised our raw residuals after removing the mean component of the model $\mathbf{X}\hat{\boldsymbol{\beta}}$, since $\hat{\mathrm{var}}[\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\Sigma}}$. This can be done by computing the Cholesky decomposition of $\hat{\boldsymbol{\Sigma}}$, where we find a lower triangular matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^T = \hat{\boldsymbol{\Sigma}}$. Then $\hat{\mathrm{var}}[\mathbf{L}^{-1}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})] = \mathbf{L}^{-1}\hat{\boldsymbol{\Sigma}}\{\mathbf{L}^{-1}\}^T = \mathbf{I}$ (the identity matrix), and so $\mathbf{L}^{-1}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})$ should be independent standard normal residuals. Using maximum likelihood estimation, we obtained the parameter estimates $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (108.27, -1.414, 1.144, 0.0068)^T$ and $\hat{\boldsymbol{\phi}}_{\mathrm{ML}} = (5.087, 3.282, 0.392)^T$, which results in the diagnostic plots presented in Figure 4.4.

We observe that the residuals are predominantly centred around zero and exhibit no patterns with respect to the fitted values. However, a non-negligible proportion of the residuals are large in magnitude relative to what is expected from the standard normal distribution, which is also clear from the quantile plot. This suggests that the residuals are heavy-tailed, and the model could be improved. Potential solutions include introducing more covariates such as the elevation of the land surface stations, trying different

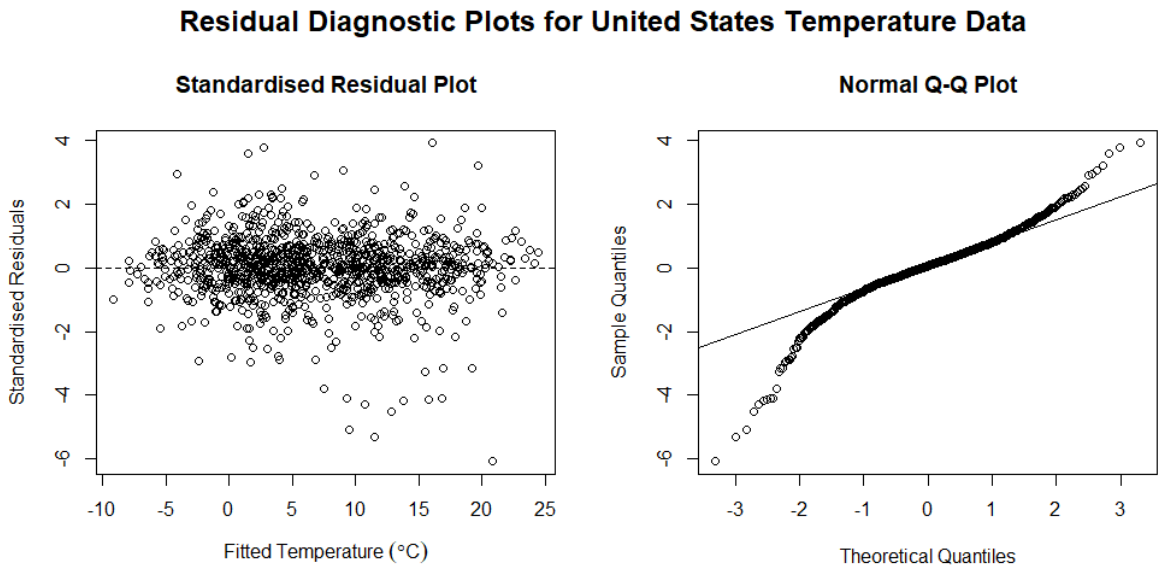

Figure 4.4: Standardised residual diagnostic plots for the isotropic exponential covariance spatial regression model. The plots suggests that the residuals are heavy-tailed.

covariance structures, or fitting a non-Gaussian model. Nevertheless, given that our primary focus in this chapter is to compare various composite likelihoods under the exponential covariance structure in two dimensions, we have proceeded with this model.

In our comparison of composite likelihoods, we have investigated both the statistical and computational performance. Whilst running the algorithm to iteratively estimate the parameters and calculate 95% Wald confidence intervals, we have recorded the mean time taken for a single iteration of the generalised least squares and Fisher scoring steps, and also the calculation of $G_{\phi}(\hat{\boldsymbol{\theta}}_{\mathrm{CL}})$. We have narrowed down our analysis to the parameters $\phi = (\sigma^2, \alpha, \tau^2)^T$ as this is of greater interest here, although (4.3) and (4.5) can be implemented to obtain variance estimates for $\hat{\boldsymbol{\beta}}_{\mathrm{CL}}$ if desired.

For the composite marginal blockwise likelihood, we tested a range of average block sizes from $W = 2$ to $W = N = 1052$ (i.e. the full likelihood). In Figure 4.5, we see that the maximum composite likelihood estimates for each parameter are well within the 95% confidence interval for the corresponding maximum likelihood estimate and are hence reasonable. The interval widths are of a similar length for σ^2, but the intervals for lower average block sizes are more notably longer for α and τ^2, with $W = 2$ having especially poor precision. This matches our findings in Section 3.4.3 where the maximum composite

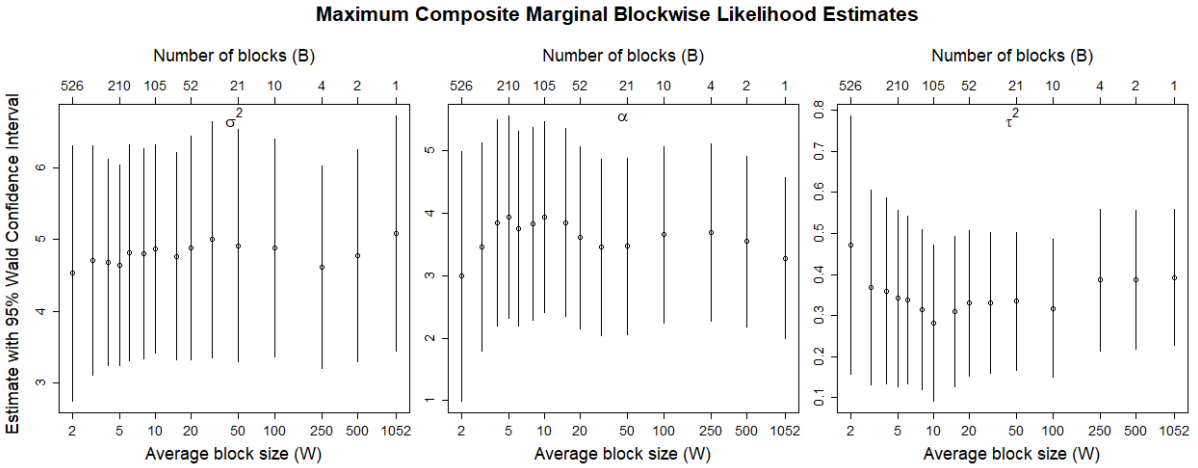

Figure 4.5: Maximum composite marginal blockwise likelihood estimation for the maximum temperature spatial regression model with covariance parameters $\boldsymbol{\theta} = (\sigma^2, \alpha, \tau^2)^T$. Estimates and Wald confidence intervals for different block sizes are comparable to the maximum likelihood estimate ($W = 1052$), but interval widths generally decrease as W increases.

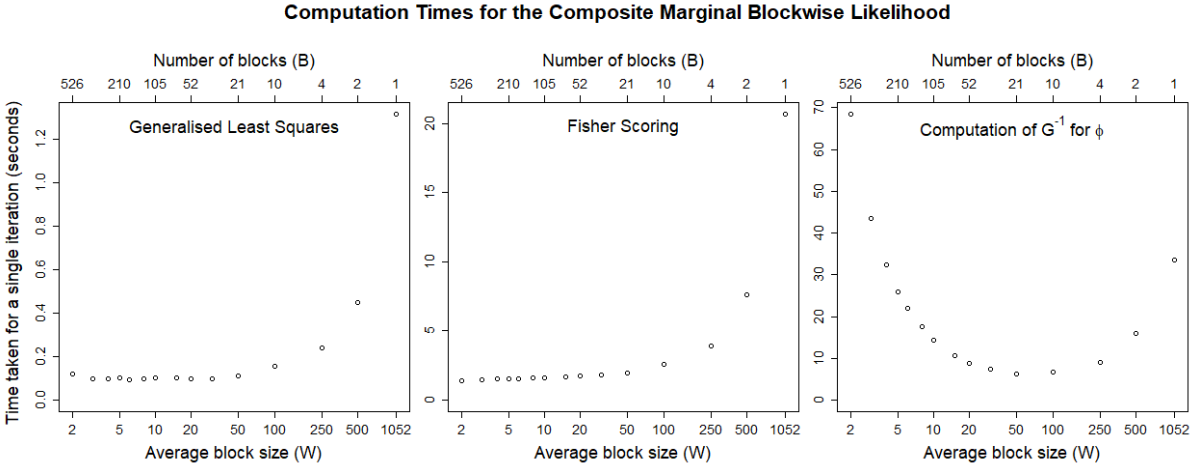

Figure 4.6: Runtime of various computations for the composite marginal blockwise likelihood. Left and middle: the iterative estimation procedure is fastest for small block sizes. Right: there is a U-shaped relationship in the computation time of the sandwich covariance matrix.

marginal blockwise likelihood estimator for σ^2 has a consistently high relative efficiency, whereas the efficiency of α is often poor for low block sizes.

In terms of computation times, performing the iterative estimation scheme on composite marginal blockwise likelihoods is naturally significantly faster than the full likelihood, as seen in Figure 4.6. Smaller block sizes allow us to avoid the $O(N^3)$ complexity computations of determinants and inverses for large matrices, which are a part of the generalised least squares and Fisher scoring steps in (4.2). The number of iterations until convergence for each block size were also very similar. On the other hand, small block

Likelihood	K	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\tau}^2$
Full		5.087 (3.443, 6.731)	3.282 (1.992, 4.572)	0.392 (0.227, 0.557)
K-Nearest	1	4.871 (3.415, 6.327)	4.314 (2.178, 6.449)	0.299 (-0.063, 0.660)
	2	4.851 (3.360, 6.341)	4.136 (2.098, 6.174)	0.323 (0.037, 0.609)
	3	5.402 (3.335, 7.468)	3.829 (1.698, 5.961)	0.288 (0.035, 0.540)
K-Sequential	1	4.820 (3.426, 6.214)	4.236 (2.497, 5.975)	0.215 (-0.050, 0.481)
	2	4.695 (3.196, 6.194)	3.696 (2.137, 5.254)	0.360 (0.154, 0.566)
	3	4.638 (3.172, 6.104)	3.719 (2.165, 5.273)	0.386 (0.187, 0.585)

Table 4.1: Estimates and 95% Wald confidence intervals for various choices of composite conditional likelihood compared to the full likelihood.

sizes are disadvantageous when estimating the variance using $\mathbf{G}_\phi(\hat{\boldsymbol{\theta}})^{-1}$, which is predominantly due to the form of $\mathbf{J}_\phi(\hat{\boldsymbol{\theta}})$ in (4.6). In this expression, we have a double summation based on the number of blocks, so increasing the average block size can greatly reduce the number of traces of eight-matrix products that need to be computed. Nevertheless, this computation has a complexity of $O(N^2)$ for a fixed block size W, which is preferred over $O(N^3)$ under the full likelihood when N is large.

The composite conditional K-nearest neighbours likelihood and K-sequential neighbours likelihood were also tested for different values of K. When compared to the maximum likelihood estimates, we can see from Table 4.1 that when $K = 1$, both composite conditional likelihoods seem to estimate σ^2 well but have point estimates that are not as close for α and τ^2. This is a potential side-effect of their composition only involving univariate and bivariate densities, making it difficult to capture information about all three parameters. Comparatively, higher values of K for the composite conditional K-sequential neighbours likelihood appear to improve the inference on α and τ^2 in particular. The differences in estimation and inference between $K = 2$ and $K = 3$ are also quite small for this composite likelihood, suggesting that a small value of K may be suitable for faster computation without foregoing much precision. For these values of K, the K-sequential neighbours likelihood appears to have similar length confidence intervals to the composite marginal blockwise likelihood with small average block sizes from Figure 4.5.

A few issues were noted with the iterative algorithm when applied to these composite conditional likelihoods. In the case where $K = 4$ for the K-nearest neighbours likelihood, the algorithm failed to converge as the parameter α reached negative values regardless of the starting point. This is potentially a structural

Likelihood	K	Generalised Least Squares	Fisher Scoring	$\mathbf{G}_{\phi}(\hat{\boldsymbol{\theta}})^{-1}$
Full		1.30	20.8	34
K-Nearest	1	0.30	3.9	865
	2	0.48	7.1	1269
	3	0.68	10.4	1697

Table 4.2: Runtime (in seconds) of various computations for the composite conditional K-nearest neighbours likelihood compared to the full likelihood. The composite conditional K-sequential neighbours likelihood has been omitted as it has similar computation times to the K-nearest neighbours likelihood for each K.

problem where the composite likelihood exhibits instability as K increases. Furthermore, the confidence intervals when $K = 3$ already indicate this as a possibility considering that they are somewhat longer than when $K = 2$, especially for σ^2. The K-sequential neighbours likelihood with $K = 1$ also had an issue where the updated values fluctuated above and below the maximum composite likelihood estimate with very slow convergence. This highlights that for the K-sequential neighbours likelihood, structural instability can occur if K is too low.

With regards to computation, the composite conditional likelihoods allow us to obtain point estimates faster than maximum likelihood estimation, but variance estimation using (4.6) is very costly. Our general representation of any composite log-likelihood as a linear combination of marginal densities allows us to write the K-nearest neighbours likelihood in terms of N blocks of size $K+1$ and N blocks of size K, while the K-sequential neighbours likelihood involves $N-K$ blocks of size $K+1$ and $N-K-1$ blocks of size K. For small values of K, these quantities are comparable, leading to similar computational performance. However, this means that the computation of $\mathbf{J}_{\phi}(\hat{\boldsymbol{\theta}})$ involves a double summation of roughly $(2N)^2$ terms, which is far worse than the $(N/W)^2$ terms in a composite marginal blockwise likelihood. In practice, however, we would exploit the structure of our selected composite likelihood to implement it more efficiently rather than rely on this general expression to estimate the variance.

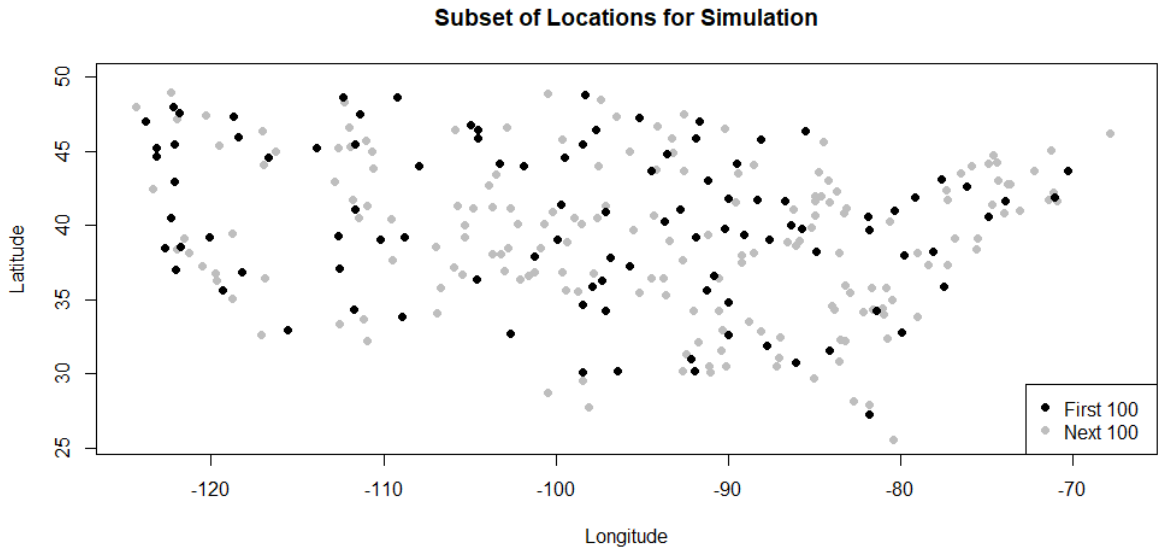

Figure 4.7: Random subset of locations used for simulations. Black points are those in the sample when $N = 100$ and grey points are also included when $N = 200$. This sampling scheme is most similar to infill.

4.5 Simulation Study

In order further investigate the statistical properties of the proposed point and variance estimators in the two-dimensional setting with irregularly-spaced observations, we performed a data-motivated simulation study. This involved taking a random subset of the $N = 1052$ observation locations, as shown in Figure 4.7, and repeatedly generating samples from these locations according to a true model. We have considered both a true model without the nugget effect, which has allowed us to make more direct comparisons with our results from Chapter 3, as well as the more realistic situation of a true model with a nugget effect included. In both situations, we have simulated multivariate normal data with zero mean and covariance structure $\Sigma_{0,ij} = \tau_0^2 I(i = j) + \sigma_0^2 \exp(-\alpha_0 \|\mathbf{s}_i - \mathbf{s}_j\|)$, where $\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$ and $\tau_0^2 = 0$ under the nuggetless model and $\tau_0^2 = 0.4$ with the nugget effect. These parameter values are the maximum likelihood estimates from our maximum temperature data, rounded to one decimal place. Our primary interest was to investigate the bias, relative efficiency and coverage probability of the Wald confidence interval for the different choices of composite likelihood.

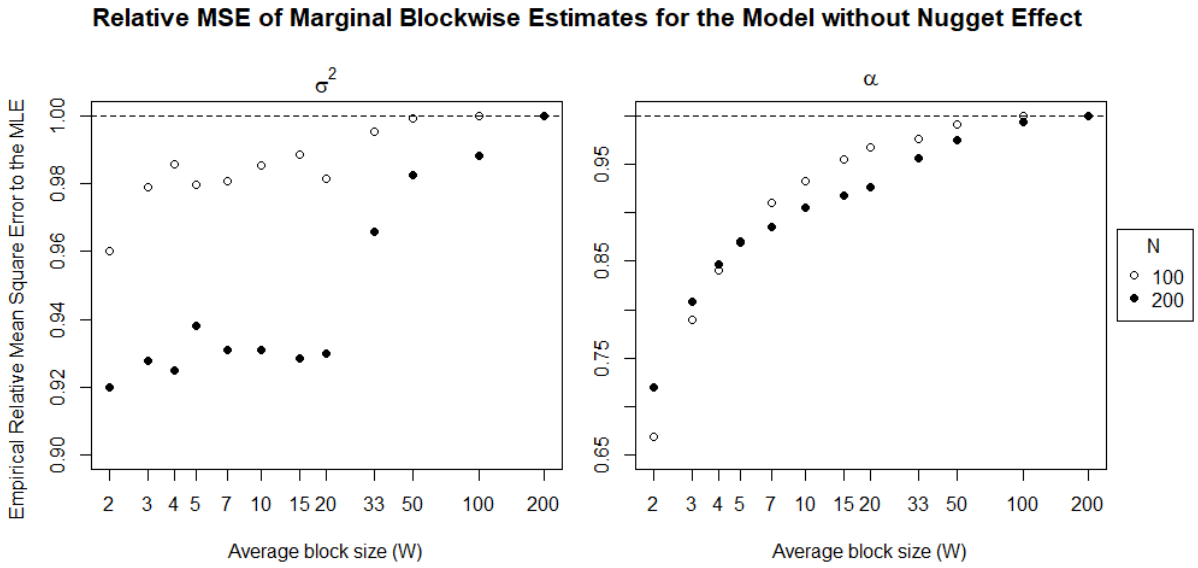

Figure 4.8: Empirical relative efficiency of maximum composite marginal blockwise likelihood estimation for $\sigma_0^2 = 5.1$ and $\alpha_0 = 3.3$ (5000 simulations).

4.5.1 Model without Nugget Effect

The empirical bias of each of the maximum composite likelihood estimators tested was similar to that of the maximum likelihood estimator. As an example, for a sample size of 100, the bias of $\hat{\sigma}^2_{\text{ML}}$ was around 0.013 and the bias of $\hat{\alpha}_{\text{ML}}$ was around 0.224. In comparison, the bias of the maximum composite marginal blockwise likelihood estimator with $W = 2$ was 0.006 for $\hat{\sigma}^2_{\text{CL}}$ and 0.259 for $\hat{\alpha}_{\text{CL}}$, which are also positive. This is due to the sampling distributions of the estimators for both parameters being right-skewed; though the positive bias is much smaller for σ^2 than for α. Regardless, the bias of these estimators was considerably smaller than their variance, and will therefore have little impact on our analysis of relative efficiency.

For the composite marginal blockwise likelihood, the relative efficiencies in the two-dimensional setting exhibited similarities to our findings for the one-dimensional case in Section 3.4.3. From Figure 4.8, we see that the relative efficiency generally increases with the average block size; and α in particular exhibits the same plateauing curve as Figure 3.3 for weaker dependence between nearby observations. Additionally, we can see that for low block sizes, the relative efficiency is higher for $N = 200$ than

Likelihood	K	σ^2		α	
		$N = 100$	$N = 200$	$N = 100$	$N = 200$
K-Nearest	2	0.670	0.451	0.865	0.641
K-Sequential	1	0.975	0.901	0.906	0.823
	2	0.972	0.906	0.910	0.848
	3	0.977	0.913	0.918	0.884

Table 4.3: Empirical relative efficiency of various choices of composite conditional likelihood for $\sigma_0^2 = 5.1$ and $\alpha_0 = 3.3$ (1000 simulations).

$N = 100$. This is consistent with the idea that the further observations taken for the sample size of 200 in Figure 4.7 aligns with the infill asymptotic framework. For σ^2, we also notice similarities to Figure 3.3, such as the relative efficiency being quite high at above 0.9, and reaching a lower minimum under infill.

On the other hand, the efficiency of the maximum composite conditional 2-nearest neighbours likelihood show differences from the asymptotic efficiency in one dimension. In Table 4.3, we observe that the efficiency of σ^2 is lower than that of α. However, the expanding domain asymptotic relative efficiencies in (3.23) satisfy $\text{ARE}(\hat{\sigma}^2_{\text{CL}}, \hat{\sigma}^2_{\text{ML}}) \geq \text{ARE}(\hat{\alpha}_{\text{CL}}, \hat{\alpha}^2_{\text{ML}})$, as illustrated in Figure 3.2. This suggests that the behaviour of the composite conditional K-nearest neighbours likelihood may differ in the two-dimensional case, although this is of course subject to simulation design. In fact, when estimates for $K = 3$ were simulated, $\hat{\sigma}^2_{\text{CL}}$ attained values in the thousands on several occasions, while $\hat{\alpha}_{\text{CL}}$ was still quite stable, indicating an inherent structural issue with this composite likelihood. Additionally, by comparing $N = 100$ to $N = 200$ in Table 4.3, we still see the issue where efficiency decreases under infill.

The efficiency of the estimators from the K-sequential neighbours likelihood are quite high even for small values of K. Table 4.3 suggests that the efficiency increases slowly with K, which is expected due to the improvement in approximation to the full likelihood. However, even when $K = 1$, we attain efficiencies of above 0.8 for both parameters, which has comparable performance to the composite marginal blockwise likelihood in Figure 4.8 with small block sizes.

In terms of coverage probabilities, all choices of composite likelihood achieve similar actual coverage

Likelihood		σ^2		α	
		$N = 100$	$N = 200$	$N = 100$	$N = 200$
Full		0.905	0.932	0.969	0.964
Marginal Blockwise	$W = 2$	0.910	0.928	0.961	0.965
	5	0.908	0.932	0.972	0.968
	10	0.909	0.934	0.971	0.964
Conditional K-Nearest	$K = 2$	0.917	-	0.959	-
Conditional K-Sequential	$K = 1$	0.906	-	0.968	-
	2	0.907	-	0.969	-
	3	0.908	-	0.968	-

Table 4.4: Empirical coverage probabilities of the 95% Wald confidence interval for various choices of composite likelihood, with variance estimated using $\mathbf{G}(\hat{\sigma}^2, \hat{\alpha})^{-1}$ ($\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$, 1000 simulations). Empty entries are due to the required computation times for obtaining these values being impractical.

Likelihood		σ^2	α	τ^2	$\hat{P}(\hat{\alpha} = 0)$	$\hat{P}(\hat{\tau}^2 = 0)$
Full		-0.120	0.075	0.125	0.000	0.357
Marginal Blockwise	$W = 2$	-0.273	-0.291	0.270	0.092	0.405
	5	-0.193	-0.051	0.193	0.004	0.365
	10	-0.162	-0.012	0.167	0.003	0.365
Conditional K-Nearest	$K = 2$	-0.077	-0.184	0.192	0.012	0.415
Conditional K-Sequential	$K = 1$	-0.278	-0.144	0.249	0.021	0.403
	2	-0.163	0.054	0.144	0.007	0.389
	3	-0.141	0.089	0.128	0.003	0.369

Table 4.5: Empirical bias and proportion of zero estimates for various choices of composite likelihood ($\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$, $\tau_0^2 = 0.4$, $N = 100$, 1000 simulations).

to the full likelihood. From Table 4.4, we see that using the Wald confidence interval with estimated variance $\mathbf{G}(\hat{\sigma}^2, \hat{\alpha})^{-1}$ provides undercoverage of σ^2 and overcoverage of α in this situation; though this improves as N increases since these estimators approach their respective normal sampling distributions.

4.5.2 Model with Nugget Effect

The addition of the small measurement error of $\tau_0^2 = 0.4$ into the Gaussian exponential covariance model introduced notable issues with estimation. Primarily, for a relatively small sample size such as $N = 100$, even the full likelihood sometimes had its maximum on the boundary $\tau^2 = 0$. Hence, we had to use a general purpose optimiser with box constraints to obtain our estimates rather than Fisher scoring. For such a small sample size and only three parameters, this had little effect on computation time.

From Table 4.5, we can see that the nugget effect has a significant effect on the bias of the estimators.

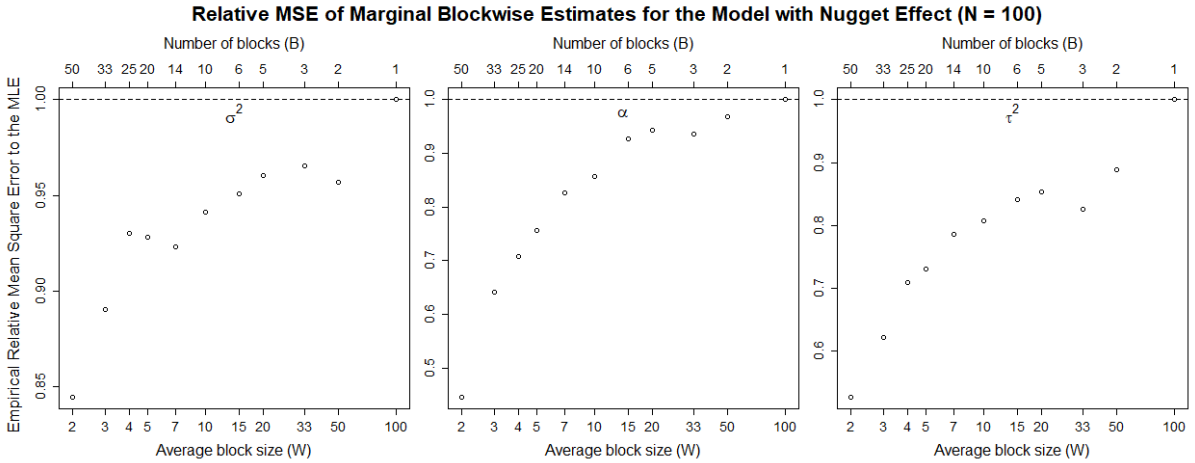

Figure 4.9: Empirical relative efficiency of maximum composite marginal blockwise likelihood estimation for $\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$ and $\tau_0^2 = 0.4$ (1000 simulations).

Likelihood	K	σ^2	α	τ^2
K-Nearest	2	0.759	0.645	0.696
K-Sequential	1	0.829	0.647	0.545
	2	0.944	0.803	0.821
	3	0.993	0.852	0.914

Table 4.6: Empirical relative efficiency for various choices of composite conditional likelihood for $\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$ and $\tau_0^2 = 0.4$ ($N = 100$; 1000 simulations).

The bias of $\hat{\tau}^2$ for all likelihoods is quite large relative to its true value of 0.4, which is attributable to the large proportion of estimates that ended up on the zero boundary. As a side-effect of the nugget effect being a competing variance component to σ^2, the bias of $\hat{\sigma}^2$ is now negative, which is due to the negative correlation between $\hat{\sigma}^2$ and $\hat{\tau}^2$.

Compared to the nuggetless case, we find that the bias is heavily affected by the choice of composite likelihood. In particular, the bias of estimators based on composite likelihoods involving very small blocks are particularly far away from those of the full likelihood, to the extent of having a negative bias for the estimate of $\hat{\alpha}_{\mathrm{CL}}$. This is likely due to the small blocks, especially those of size 1 and 2, being inadequate for capturing information about all three parameters. However, we can see that this quickly improves as W increases for the composite marginal blockwise likelihood and K increases for the composite conditional K-sequential likelihood, where we also see far less estimates of α on the boundary.

Likelihood		σ^2	α	τ^2
Full		0.963	0.942	0.963
Marginal Blockwise	$W = 2$	0.991	0.959	1.000
	5	0.971	0.924	0.947
	10	0.959	0.929	0.950
Conditional K-Nearest	$K = 2$	0.977	0.923	0.961
Conditional K-Sequential	$K = 1$	0.977	0.916	0.938
	2	0.969	0.926	0.951
	3	0.962	0.943	0.958

Table 4.7: Empirical coverage probabilities of the 95% Wald confidence interval for various choices of composite likelihood, with variance estimated using $\mathbf{G}(\hat{\sigma}^2, \hat{\alpha}, \hat{\tau}^2)^{-1}$ ($\sigma_0^2 = 5.1$, $\alpha_0 = 3.3$, $\tau_0^2 = 0.4$, $N = 100$; 1000 simulations).

The presence of the nugget effect also adversely affected the relative efficiency of the maximum composite likelihood estimators. Compared to Figure 4.8 and Table 4.3, we can see in Figure 4.9 and Table 4.6 that the relative efficiencies are generally at lower levels than in the nuggetless case. Thus, it is advised that when there are more parameters to estimate in the spatial covariance model that a composite likelihood involving larger block sizes is chosen.

Finally, Table 4.7 shows that all choices of composite likelihood provide coverage near the nominal 95% level except for the marginal blockwise likelihood with $W = 2$. In this case, we have overcoverage for all three parameters, suggesting that the normal approximation of the sampling distribution is poor. This can be traced back to Table 4.5, where this likelihood has the highest proportion of zero estimates for α at 0.092.

Chapter 5

Conclusion

The primary focus of this thesis was to investigate the statistical and computational performance of maximum composite likelihood estimation as an alternative to maximum likelihood estimation. This is of particular importance in a geostatistical model due to the dependence structure often making the full likelihood computationally expensive to evaluate for large sample sizes. From both a theoretical and computational standpoint, we analysed the relative efficiency for various choices of composite likelihood under the Gaussian isotropic exponential covariance model. Out of the composite likelihoods investigated, we would recommend using the composite marginal blockwise likelihood and composite conditional K-sequential neighbours likelihood as they had relatively high efficiencies in the cases considered. Since these two composite likelihoods are structured in such a way as to approximate the full likelihood, as opposed to the composite conditional K-nearest neighbours likelihood, there is some flexibility in the choice of block size based on a trade-off between computation time and relative efficiency.

5.1 Main Contributions

We have made several contributions to the literature on composite likelihood, particularly in a Gaussian geostatistical setting. In the one-dimensional exponential covariance model, we derived the exact form of the sandwich covariance matrix for the composite conditional 2-nearest neighbours likelihood and composite marginal blockwise likelihood. This was done in a setting with equally-spaced observation locations that was flexible enough to consider all three asymptotic frameworks via the expanding

domain parameter D and infill parameter F, thereby allowing us to explore the implications of each framework. The derivation in both cases involved combining all of the individual densities together to form a composition matrix $\mathbf{M}$ whose structure was additively decomposed. By doing this, we overcame the hurdle of deriving $\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}[\mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{y})\mathrm{sc}_{\mathrm{C}}(\boldsymbol{\theta};\mathbf{y})^T]$, which is frequently problematic due to the presence of four-order moments.

Due to our theoretical derivations, we discovered new insights into the statistical performance of these composite likelihoods. In an expanding domain framework, the composite conditional 2-nearest neighbours likelihood is fully efficient for both σ^2 and α when the data are uncorrelated ($\alpha_0 = \infty$), with asymptotic relative efficiency decreasing to zero as the strength of dependence between adjacent observations increases ($\alpha_0 \to 0$ or $F \to \infty$). Hence, we would only recommend using this composite likelihood for data that are spaced quite far apart. For the composite marginal blockwise likelihood, we discovered an interesting relationship between block size and the expanding domain asymptotic relative efficiency. These results suggest that choosing small block sizes can sometimes outperform larger block sizes, which we attribute to a "between-blocks" effect that helps to capture information about σ^2. Additionally, under the hybrid framework, both σ^2 and α are fully efficient for any block size of at least two.

In our data-motivated study of the two-dimensional exponential covariance process with irregularly-spaced observations, our main contribution was the development of a general framework that allows for the analysis of both composite marginal and composite conditional likelihood functions. This is possible due to the key observation that any composite log-likelihood can be expressed as a linear combination of marginal log-densities. Specifically, we generalised the work of Mardia and Marshall (1984) to the composite likelihood setting in order to obtain an iterative estimation algorithm and exact expression for the sandwich covariance matrix under a Gaussian spatial regression model; both of which can be implemented computationally.

Our simulation study in the two-dimensional setting allowed us to make comparisons to our theoretical

results in the one-dimensional setting without a nugget effect. We observed that the behaviour of the composite marginal blockwise likelihood is similar with respect to the level of relative efficiency and effect of infill. However, the composite conditional 2-nearest neighbours likelihood may behave differently in the two-dimensional setting as the asymptotic relative efficiency of α appears to be higher than that of σ^2, whereas it is the opposite in one dimension. We also explored the implications of a nugget effect on the bias and efficiency of various maximum composite likelihood estimators, as well as coverage probabilities of Wald confidence intervals.

5.2 Further Research

A number of other cases may be considered when it comes to exact derivations of the sandwich covariance matrix. For instance, there is some ambiguity in how observations are grouped in the composite marginal blockwise likelihood, and the sequence of observations in the composite conditional K-sequential neighbours likelihood. In the case of the marginal blockwise likelihood, a block could be formed by starting from the observation indexed by $q \in \{1,2,...,W\}$ and taking every B-th observation thereafter on the number line. Given the poor performance of the composite conditional K-nearest neighbours likelihood under infill, we could also consider a composite likelihood that only includes every second conditional density; that is, $\mathcal{L}_C(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{\lceil \frac{N-1}{2} \rceil} f(y(s_{2i})|y(s_{2i-1}),y(s_{2i+1});\boldsymbol{\theta})$. This is more in line with the original construction of Besag (1974). Finally, the composite conditional and composite marginal pairwise likelihoods are also of interest; though since these likelihood functions incorporate all pairs of observations, a potential complication is that the composition matrix will not be sparse.

Theoretical results for maximum composite likelihood estimation could also be obtained in a two-dimensional exponential covariance setting on an equally-spaced lattice. Results under the full likelihood case where $C(\mathbf{h}) = \sigma^2 \exp(-\alpha_1|h_1| - \alpha_2|h_2|))$ for a separation vector $\mathbf{h} = (h_1,h_2)^T$ are available due to Ying (1993). In particular, the separation of the covariance structure by direction allows the

covariance matrix to be expressed as a Kronecker product of exponential covariance matrices from the one-dimensional case. This case is especially notable due to the parameters being consistent and asymptotically normal under infill.

The two-dimensional case also has implications in the context of spatio-temporal data, where the temporal dimension has a naturally expanding domain. Note that the aforementioned covariance structure can be slightly restructured to the separable and stationary spatio-temporal covariance structure $C(h,t) = C(h)C(t) = \sigma^2 \exp(-\alpha_1|h|)\exp(-\alpha_2|t|)$. Thus, it would also be interesting to consider whether we have consistency in α_1 if the spatial lattice scheme remains fixed and finite at all time points, and only the temporal domain expands.

It would also be worth considering other stationary covariance models that are used in practice, such as the Matérn covariance structure or non-separable models. In these cases, an important consideration would be whether the inverse of the covariance matrix is attainable in a closed-form, or at least a sparse approximation to it.

Appendix A

Detailed derivation of the trace of a four-matrix product

We will present a detailed derivation of $\text{tr}(\mathbf{M\Sigma M\Sigma})$ from (3.20) in Section 3.3.2, where

$$\begin{aligned}\mathbf{M} &= \frac{1}{1-\rho^2}\left[\Big(1+\frac{\rho^2}{1+\rho^2}\Big)\mathbf{I}+2\rho^2\mathbf{A}+\Big(-\rho^2+\frac{\rho^2}{1+\rho^2}\Big)\mathbf{B}-2\rho\mathbf{C}+\frac{\rho^2}{1+\rho^2}\mathbf{D}\right]\\ &= \frac{1}{1-\rho^4}[(1+2\rho^2)\mathbf{I}+2(\rho^2+\rho^4)\mathbf{A}-\rho^4\mathbf{B}-2(\rho+\rho^3)\mathbf{C}+\rho^2\mathbf{D}],\end{aligned}$$

as per (3.13). Using the individual traces of four-matrix products from (3.19), we have that

$$\begin{aligned}\text{tr}(\mathbf{M\Sigma M\Sigma}) = \frac{1}{(1-\rho^4)^2}[&(1+4\rho^2+4\rho^4)\text{tr}(\mathbf{I\Sigma I\Sigma})+4\rho^4(1+2\rho^2+\rho^4)\text{tr}(\mathbf{A\Sigma A\Sigma})+\rho^8\text{tr}(\mathbf{B\Sigma B\Sigma})\\ &+4\rho^2(1+2\rho^2+\rho^4)\text{tr}(\mathbf{C\Sigma C\Sigma})+\rho^4\text{tr}(\mathbf{D\Sigma D\Sigma})\\ &+4\rho^2(1+3\rho^2+2\rho^4)\text{tr}(\mathbf{I\Sigma A\Sigma})-2\rho^4(1+2\rho^2)\text{tr}(\mathbf{I\Sigma B\Sigma})\\ &-4\rho(1+3\rho^2+2\rho^4)\text{tr}(\mathbf{I\Sigma C\Sigma})+2\rho^2(1+2\rho^2)\text{tr}(\mathbf{I\Sigma D\Sigma})\\ &-4\rho^6(1+\rho^2)\text{tr}(\mathbf{A\Sigma B\Sigma})-8\rho^3(1+2\rho^2+\rho^4)\text{tr}(\mathbf{A\Sigma C\Sigma})\\ &+4\rho^4(1+\rho^2)\text{tr}(\mathbf{A\Sigma D\Sigma})+4\rho^5(1+\rho^2)\text{tr}(\mathbf{B\Sigma C\Sigma})\\ &-2\rho^6\text{tr}(\mathbf{B\Sigma D\Sigma})-4\rho^3(1+\rho^2)\text{tr}(\mathbf{C\Sigma D\Sigma})]\\ = \frac{\sigma^4}{(1-\rho^4)^2}&[Q_1+Q_2],\end{aligned}$$

where

$$
\begin{aligned}
Q_1 =&(1+4\rho^2+4\rho^4)DF+4\rho^4(1+2\rho^2+\rho^4)(DF-2)+\rho^8(DF-4)\\
&+8\rho^2(1+3\rho^2+3\rho^4+\rho^6)(DF-1)+2\rho^4(1+\rho^4)(DF-2)+4\rho^6(1+\rho^2)(DF-3)\\
&+4\rho^2(1+3\rho^2+2\rho^4)(DF-2)-2\rho^4(1+4\rho^2+4\rho^4)(DF-4)\\
&-16\rho^2(1+3\rho^2+2\rho^4)(DF-1)+4\rho^4(1+2\rho^2)(DF-2)\\
&-4\rho^6(1+\rho^2)(DF-4)-32\rho^4(1+2\rho^2+\rho^4)(DF-2)\\
&+8\rho^6(1+\rho^2)(DF-2)+16\rho^6(1+\rho^2)(DF-4)\\
&-4\rho^8(DF-4)-16\rho^4(1+2\rho^2+\rho^4)(DF-2)\\
=&(1+4\rho^2+4\rho^4)DF+(-8\rho^2-24\rho^4-8\rho^6+8\rho^8)(DF-1)\\
&+(4\rho^2-26\rho^4-64\rho^6-34\rho^8)(DF-2)+(4\rho^6+4\rho^8)(DF-3)+(-2\rho^4+4\rho^6+\rho^8)(DF-4)\\
=&(1-48\rho^4-64\rho^6-21\rho^8)DF+84\rho^4+108\rho^6+44\rho^8.
\end{aligned}
$$

By repeatedly applying $u_n=\frac{u_{n+1}}{\rho^2}-n$, we also have

$$
\begin{aligned}
Q_2 =& 2[(1+4\rho^2+4\rho^4)u_{DF}+4\rho^4(1+2\rho^2+\rho^4)u_{DF-2}+\rho^8u_{DF-4}\\
&+16\rho^2(1+2\rho^2+\rho^4)u_{DF-1}+4\rho^6u_{DF-3}\\
&+4\rho^2(1+3\rho^2+2\rho^4)u_{DF-1}-2\rho^6(1+2\rho^2)u_{DF-3}\\
&-8\rho^2(1+3\rho^2+2\rho^4)u_{DF-1}+4\rho^2(1+2\rho^2)u_{DF-1}\\
&-4\rho^6(1+\rho^2)u_{DF-3}-16\rho^4(1+2\rho^2+\rho^4)u_{DF-2}\\
&+8\rho^4(1+\rho^2)u_{DF-2}+8\rho^6(1+\rho^2)u_{DF-3}\\
&-4\rho^6u_{DF-3}-16\rho^4(1+\rho^2)u_{DF-2}]\\
=& 2[(1+4\rho^2+4\rho^4)u_{DF}+2\rho^2(8+14\rho^2+4\rho^4)u_{DF-1}\\
&-4\rho^4(5+8\rho^2+3\rho^4)u_{DF-2}+2\rho^6u_{DF-3}+\rho^8u_{DF-4}]
\end{aligned}
$$

$$
\begin{aligned}
&= 2[(1+4\rho^2+4\rho^4)u_{DF} + 2\rho^2(8+14\rho^2+4\rho^4)u_{DF-1} \\
&\quad - 4\rho^4(5+8\rho^2+3\rho^4)u_{DF-2} + 3\rho^6 u_{DF-3} - \rho^8(DF-4)] \\
&= 2[(1+4\rho^2+4\rho^4)u_{DF} + 2\rho^2(8+14\rho^2+4\rho^4)u_{DF-1} \\
&\quad - \rho^4(17+32\rho^2+12\rho^4)u_{DF-2} - 3\rho^6(DF-3) - \rho^8(DF-4)] \\
&= 2[(1+4\rho^2+4\rho^4)u_{DF} - \rho^2(1+4\rho^2+4\rho^4)u_{DF-1} \\
&\quad + \rho^4(17+32\rho^2+12\rho^4)(DF-2) - 3\rho^6(DF-3) - \rho^8(DF-4)] \\
&= 2[\rho^2(1+4\rho^2+4\rho^4)(DF-1) + \rho^4(17+32\rho^2+12\rho^4)(DF-2) - 3\rho^6(DF-3) - \rho^8(DF-4)] \\
&= 2[(\rho^2+21\rho^4+33\rho^6+11\rho^8)DF - \rho^2 - 38\rho^4 - 59\rho^6 - 20\rho^8].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathrm{tr}(\mathbf{M\Sigma M\Sigma}) &= \frac{\sigma^4}{(1-\rho^4)^2}[Q_1 + Q_2] \\
&= \frac{\sigma^4}{(1-\rho^4)^2}[(1+2\rho^2-6\rho^4+2\rho^6+\rho^8)DF - 2\rho^2 + 8\rho^4 - 10\rho^6 + 4\rho^8] \\
&= \frac{\sigma^4}{(1-\rho^4)^2}[(1-\rho^2)^2(1+4\rho^2+\rho^4)DF - 2\rho^2(1-\rho^2)^2(1-2\rho^2)] \\
&= \frac{\sigma^4}{(1+\rho^2)^2}[(1+4\rho^2+\rho^4)DF - 2\rho^2 + 4\rho^4],
\end{aligned}
$$

as required.

This procedure can be repeated to obtain expressions for $\mathrm{tr}(\mathbf{M\Sigma M'\Sigma})$ and $\mathrm{tr}(\mathbf{M'\Sigma M'\Sigma})$. Similarly, we can use this method to find the traces of the four-matrix products in (3.34) for the composite marginal blockwise likelihood.

Bibliography

Angelini, M. E., & Heuvelink, G. B. M. (2018). Including spatial correlation in structural equation modelling of soil properties. *Spatial Statistics*, *25*, 35–51.

Bachoc, F., Bevilacqua, M., & Velandia, D. (2018). Composite likelihood estimation for a Gaussian process under fixed domain asymptotics. *ArXiv e-prints*. arXiv: 1807.08988 [math.ST]

Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika*, *40*, 12–19.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, *36*, 192–236.

Bevilacqua, M., & Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing*, *25*, 877–892.

Caragea, P., & Smith, R. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, *98*, 1417–1440.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chen, H. S., Simpson, D. G., & Ying, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, *10*, 141–156.

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, *91*, 729–737.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken, NJ: John Wiley and Sons.

Davis, R. A., & Chun, Y. Y. (2011). Comments on pairwise likelihood in time series models. *Statistica Sinica*, *21*, 255–277.

Eaton, M. L. (1983). *Multivariate statistics: A vector space approach*. New York: John Wiley and Sons.

Fortin, M. J., James, P. M. A., MacKenzie, A., Melles, S. J., & Rayfield, B. (2012). Spatial statistics, spatial regression, and graph theory in ecology. *Spatial Statistics*, *1*, 100–109.

Gneiting, T., Sasvári, Z., & Schlather, M. (2001). Analogies and correspondences between variograms and covariance functions. *Advances in Applied Probability*, *33*, 617–630.

Hall, P., & Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, *99*, 399–424.

Heagerty, P. J., & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, *93*, 1099–1111.

Huang, Z., & Ferrari, D. (2017). Fast construction of efficient composite likelihood equations. *ArXiv e-prints*. arXiv: 1709.03234 [math.ST]

Hui, F. K. C., Müller, S., & Welsh, A. H. (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association*. doi:10.1080/01621459.2017.1371026

Ibragimov, I. A., & Rozanov, Y. A. (1978). *Gaussian random processes*. New York: Springer.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, *12*, 134–139.

Kac, M., Murdock, W. L., & Szegö, G. (1953). On the eigen-values of certain hermitian forms. *Journal of Rational Mechanics and Analysis*, *2*, 767–800.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, *69*, 19–27.

Lahiri, S. N., Lee, Y., & Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, *103*, 65–85.

Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, *72*, 1899–1925.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.

Lu, Z., & Tjøstheim, D. (2014). Nonparametric estimation of probability density functions for irregularly observed spatial data. *Journal of the American Statistical Association*, *109*, 1546–1564.

Mardia, K. V., Hughes, G., & Taylor, C. C. (2007). Efficiency of the pseudolikelihood for multivariate normal and von Mises distributions. *The Canadian Journal of Statistics*, *36*, 99–109.

Mardia, K. V., & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, *71*, 135–146.

Matthews, S. A., & Parker, D. M. (2013). Progress in spatial demography. *Demographic Research*, *28*, 271–312.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*, 17–23.

Nowak, G., Welsh, A. H., O'Neill, T. J., & Feng, L. B. (2017). Spatio-temporal modelling of rainfall in the Murray-Darling basin. *Journal of Hydrology*, *557*, 522–538.

Oman, S. D., & Landsman, V. (2007). Analyzing spatially distributed binary data using independent-block estimating equations. *Biometrics*, *63*, 892–900.

Peterson, T. C., & Vose, R. S. (1997). An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society*, *78*, 2837–2849.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37*, 81–89.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, *68*, 159.

Stein, M. L. (1999). *Interpolation of spatial data*. New York: Springer.

Stein, M. L., Chi, Z., & Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society Series B*, *66*, 275–296.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, *8*, 1375–1381.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*, 1–28.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

Vecchia, A. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B*, *50*, 297–312.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, *36*, 280–296.

Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, *21*, 1567–1590.

Zhang, H., & Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, *92*, 921–936.

Zheng, Y., & Zhu, J. (2012). On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. *The Indian Journal of Statistics*, *74*, 29–56.